

高等学校专业英语教材

自动化与电子信息专业英语

杨植新 主编 周 劲 副主编

孙江波 李素芬 董桂红 参编

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书主要针对自动化、电气控制和电子信息等专业的本科生阅读和翻译英文文献资料的需要而编写,选编的文献资料涵盖了电工、电子电路、电子电气设备器件、传感技术、微机原理、控制理论、计算机控制等从基础理论到实际运用的广泛内容。所有文献均出自海外原文资料。除了提供专业词汇和难句注释分析外,还为读者提供了所有课文的参考译文。绝大多数课文后还提供了课外阅读材料。除了教材、论文这些最常见的文体外,还有技术说明、产品使用及科技交流、宣传等多种文体,目的在于使读者能够多方面接触各种不同类型的英文资料。所有这些编排都十分有利于读者深入学习理解原文,提高阅读和翻译能力。

本书不仅适合电类专业本科生及研究生使用,也适合广大相关工程专业的技术人员参考。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

自动化与电子信息专业英语/杨植新主编. —北京:电子工业出版社,2009.1

高等学校专业英语教材

ISBN 978-7-121-07520-9

I. 自… II. 杨… III. 自动化—英语—高等学校—教材 IV. H31

中国版本图书馆 CIP 数据核字(2008)第 155846 号

策划编辑:凌 毅

责任编辑:谭海平

印 刷:北京市顺义兴华印刷厂

装 订:三河市双峰印刷装订有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本:787×980 1/16 印张:22.25 字数:570 千字

印 次:2009 年 1 月第 1 次印刷

印 数:4000 册 定价:29.50 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至 zlts@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010)88258888。

前 言

根据目前高等院校自动化、电气工程及其自动化及电子信息工程等专业有关课程教学大纲的要求,我们组织编写了本书。

本书共计 7 个部分 48 个单元。各单元主要由课文、专业词汇、注释和参考译文几部分组成。除第七部分外,课文均选自相关专业英文原版大学教材、专业文献。第七部分亦源自海外实际材料。课程内容覆盖相关电类专业从技术基础到专业的发展、理论和应用;文体涉及教材、论文、技术说明书及应用文等,以期让读者尽可能广泛地接触到各种与专业有关的资料。绝大部分课程附有课外阅读,有助于读者提高独立阅读能力。

本书由武汉工业学院的几位教师和辽宁工程技术大学的董桂红老师编写。他们在专业教学科研、对外交流、专业英语教学方面各有所长,有利于本书的编写。杨植新担任主编,主持本书编写大纲的制订及全书统稿工作,并编写了第二部分的第 7 单元、第四部分的第 1~4 单元和第 6 单元、第五部分的第 4 单元和第 5 单元及第七部分,第三部分由杨植新和董桂红共同编写;周劲任副主编,编写了第一部分及第六部分;孙江波编写了第二部分的第 1~6 单元和第 8、9 单元;李素芬编写了第四部分的第 5、7 单元及第五部分的第 1~3 单元。本书编写过程中,在收集资料时得到了周玉女士的热情帮助,谨表谢意。

本书提供配套的电子课件,可登录电子工业出版社的华信资源教育网 [www. hxedu. com. cn](http://www.hxedu.com.cn),注册后免费下载。

鉴于作者水平有限,加之时间仓促,书中疏漏不妥之处在所难免,敬希专家及读者赐教指正。

CONTENTS

Part 1	Fundamentals of Electric Circuits	(1)
1.1	Circuit concepts	(1)
1.1.1	Text	(1)
1.1.2	Specialized English Words	(5)
1.1.3	Notes	(6)
1.1.4	Reference Translation	(6)
1.1.5	Reading Materials	(9)
1.2	Voltage and Current Laws	(10)
1.2.1	Text	(10)
1.2.2	Specialized English Words	(14)
1.2.3	Notes	(14)
1.2.4	Reference Translation	(15)
1.2.5	Reading Materials	(17)
1.3	Network Theorems	(18)
1.3.1	Text	(18)
1.3.2	Specialized English Words	(22)
1.3.3	Notes	(22)
1.3.4	Reference Translation	(23)
1.4	First-Order Circuits	(25)
1.4.1	Text	(25)
1.4.2	Specialized English Words	(30)
1.4.3	Notes	(30)
1.4.4	Reference Translation	(31)
1.5	Sinusoidal Steady-State Circuit Analysis	(34)
1.5.1	Text	(34)
1.5.2	Specialized English Words	(38)
1.5.3	Notes	(39)
1.5.4	Reference Translation	(39)
Part 2	Electronics	(42)
2.1	Interpreting a Digital IC Datasheet	(42)
2.1.1	Text	(42)

2.1.2	Specialized English Words	(46)
2.1.3	Notes	(46)
2.1.4	Reference Translation	(47)
2.1.5	Reading Materials	(49)
2.2	Diodes and Transistors(I)	(50)
2.2.1	Text	(50)
2.2.2	Specialized English Words	(53)
2.2.3	Notes	(54)
2.2.4	Reference Translation	(55)
2.2.5	Reading Materials	(57)
2.3	Diodes and Transistors(II)	(58)
2.3.1	Text	(58)
2.3.2	Specialized English Words	(62)
2.3.3	Notes	(62)
2.3.4	Reference Translation	(63)
2.3.5	Reading Materials	(66)
2.4	The Ideal of Op-amp(I)	(67)
2.4.1	Text	(67)
2.4.2	Specialized English Words	(70)
2.4.3	Notes	(70)
2.4.4	Reference Translation	(71)
2.4.5	Reading Materials	(73)
2.5	The Ideal of Op-amp(II)	(75)
2.5.1	Text	(75)
2.5.2	Specialized English Words	(77)
2.5.3	Notes	(77)
2.5.4	Reference Translation	(78)
2.6	Boolean Algebra	(80)
2.6.1	Text	(80)
2.6.2	Specialized English Words	(82)
2.6.3	Notes	(83)
2.6.4	Reference Translation	(83)
2.7	Number System	(86)
2.7.1	Text	(86)
2.7.2	Specialized English Words	(90)

2.7.3	Notes	(90)
2.7.4	Reference Translation	(91)
2.7.5	Reading Materials	(95)
2.8	Flip-Flops and Latches	(95)
2.8.1	Text	(95)
2.8.2	Specialized English Words	(99)
2.8.3	Notes	(99)
2.8.4	Reference Translation	(100)
2.9	Programmable Logic Device	(103)
2.9.1	Text	(103)
2.9.2	Specialized English Words	(107)
2.9.3	Notes	(108)
2.9.4	Reference Translation	(109)
2.9.5	Reading Materials	(112)
Part 3	Microprocessors and Distributed Computer Control	(114)
3.1	A Brief History of the Microprocessor(I)	(114)
3.1.1	Text	(114)
3.1.2	Specialized English Words	(117)
3.1.3	Notes	(117)
3.1.4	Reference Translation	(119)
3.1.5	Reading Materials	(121)
3.2	A Brief History of the Microprocessor (II)	(122)
3.2.1	Text	(122)
3.2.2	Specialized English Words	(124)
3.2.3	Notes	(125)
3.2.4	Reference Translation	(126)
3.2.5	Reading Materials	(128)
3.3	History of the Development of the ARM Chip at Acorn	(129)
3.3.1	Text	(129)
3.3.2	Specialized English Words	(132)
3.3.3	Notes	(132)
3.3.4	Reference Translation	(133)
3.3.5	Reading Materials	(136)
3.4	Memory Organization in MCS-51 Family of Microcontrollers	(137)
3.4.1	Text	(137)

3.4.2	Specialized English Words	(142)
3.4.3	Notes	(143)
3.4.4	Reference Translation	(144)
3.4.5	Reading Materials	(147)
3.5	The Development of Computer-Based Control Systems	(148)
3.5.1	Text	(148)
3.5.2	Specialized English Words	(153)
3.5.3	Notes	(154)
3.5.4	Reference Translation	(157)
3.5.5	Reading Materials	(161)
3.6	General Concepts of Hierarchical Control	(162)
3.6.1	Text	(162)
3.6.2	Specialized English Words	(166)
3.6.3	Notes	(167)
3.6.4	Reference Translation	(169)
3.6.5	Reading Materials	(173)
Part 4	Automatic Control Theory	(175)
4.1	History of Automatic Control	(175)
4.1.1	Text	(175)
4.1.2	Specialized English words	(178)
4.1.3	Notes	(178)
4.1.4	Reference Translation	(179)
4.1.5	Reading Materials	(181)
4.2	The New Generation of Advanced Process Control	(182)
4.2.1	Text	(182)
4.2.2	Specialized English words	(186)
4.2.3	Notes	(187)
4.2.4	Reference Translation	(189)
4.2.5	Reading Materials	(191)
4.3	Feedback Fundamentals	(192)
4.3.1	Text	(192)
4.3.2	Specialized English Words	(195)
4.3.3	Notes	(196)
4.3.4	Reference Translation	(196)
4.3.5	Reading Materials	(199)

4.4	Frequency Response Methods	(199)
4.4.1	Text	(199)
4.4.2	Specialized English Words	(204)
4.4.3	Notes	(204)
4.4.4	Reference Translation	(205)
4.4.5	Reading Materials	(208)
4.5	Routh's Stability Criterion	(209)
4.5.1	Text	(209)
4.5.2	Specialized English Words	(213)
4.5.3	Notes	(214)
4.5.4	Reference Translation	(215)
4.5.5	Reading Materials	(219)
4.6	State Variable Methods	(219)
4.6.1	Text	(219)
4.6.2	Specialized English Words	(224)
4.6.3	Notes	(224)
4.6.4	Reference Translation	(225)
4.6.5	Reading Materials	(228)
4.7	Root-Locus	(229)
4.7.1	Text	(229)
4.7.2	Specialized English Words	(232)
4.7.3	Notes	(232)
4.7.4	Reference Translation	(234)
4.7.5	Reading Materials	(236)
Part 5	Sensing Technology	(238)
5.1	Sensors in Manufacturing	(238)
5.1.1	Text	(238)
5.1.2	Specialized English Words	(241)
5.1.3	Notes	(242)
5.1.4	Reference Translation	(243)
5.1.5	Reading Materials	(245)
5.2	Micro Thermocouples	(246)
5.2.1	Text	(246)
5.2.2	Specialized English Words	(249)
5.2.3	Notes	(250)

5.2.4	Reference Translation	(251)
5.2.5	Reading Materials	(254)
5.3	Pressure Microsensors	(254)
5.3.1	Text	(254)
5.3.2	Specialized English Words	(258)
5.3.3	Notes	(259)
5.3.4	Reference Translation	(260)
5.3.5	Reading Materials	(264)
5.4	QProx™ QT113 Charge-Transfer Touch Sensor (I)	(264)
5.4.1	Text	(264)
5.4.2	Specialized English Words	(268)
5.4.3	Notes	(269)
5.4.4	Reference Translation	(269)
5.4.5	Reading Materials	(272)
5.5	QProx™ QT113 Charge-Transfer Touch Sensor (II)	(273)
5.5.1	Text	(273)
5.5.2	Specialized English Words	(276)
5.5.3	Notes	(277)
5.5.4	Reference Translation	(278)
5.5.5	Reading Materials	(280)
Part 6	Electric Devices and Systems	(282)
6.1	Transformers	(282)
6.1.1	Text	(282)
6.1.2	Specialized English Words	(287)
6.1.3	Notes	(288)
6.1.4	Reference Translation	(289)
6.1.5	Reading Materials	(291)
6.2	DC Motors and AC Motors	(293)
6.2.1	Text	(293)
6.2.2	Specialized English Words	(299)
6.2.3	Notes	(300)
6.2.4	Reference Translation	(301)
6.3	Adjustable Speed Drives	(303)
6.3.1	Text	(303)
6.3.2	Specialized English Words	(306)

6.3.3	Notes	(307)
6.3.4	Reference Translation	(308)
6.3.5	Reading Materials	(310)
6.4	Power Semiconductor Devices	(312)
6.4.1	Text	(312)
6.4.2	Specialized English Words	(316)
6.4.3	Notes	(317)
6.4.4	Reference Translation	(318)
6.4.5	Reading Materials	(320)
6.5	DC Power Supply	(321)
6.5.1	Text	(321)
6.5.2	Specialized English Words	(326)
6.5.3	Notes	(326)
6.5.4	Reference Translation	(327)
6.5.5	Reading Materials	(328)
Part 7	Miscellaneous	(331)
7.1	What is Electrical Engineering? ^[1]	(331)
7.1.1	Text	(331)
7.1.2	Specialized English Words	(333)
7.1.3	Notes	(334)
7.1.4	Reference Translation	(335)
7.2	The Belief of HIMA ^[1]	(337)
7.2.1	Text	(337)
7.2.2	Specialized English Words	(338)
7.2.3	Notes	(338)
7.2.4	Reference Translation	(338)
7.2.5	Reading Materials	(339)
7.3	A Letter about Scientific Communications ^[1]	(340)
7.3.1	Text	(340)
7.3.2	Specialized English Words	(341)
7.3.3	Notes	(341)
7.3.4	Reference Translation	(341)
	Main Reference Materials	(342)

Part 1 Fundamentals of Electric Circuits

1.1 Circuit concepts

1.1.1 Text

Passive and Active Elements

An electrical device is represented by a circuit diagram or network constructed from series and parallel arrangements of two-terminal elements. The analysis of the circuit diagram predicts the performance of the actual device. A two-terminal element in general form is shown in Figure 1.1.1, with a single device represented by the rectangular symbol and two perfectly conducting leads ending at connecting points A and B ^[1]. Active elements are voltage or current sources which are able to supply energy to the network. Resistors, inductors, and capacitors are passive elements which take energy from the sources and either convert it to another form or store it in an electric or magnetic field.



Figure 1.1.1 A two-terminal element in general form.

Fig 1.1.2 illustrates seven basic circuit elements. Elements (a) and (b) are voltage sources and (c) and (d) are current sources. A voltage source that is not affected by changes in the connected circuit is an independent source, illustrated by the circle in Figure 1.1.2(a). A dependent voltage source which changes in some described manner with the conditions on the connected circuit is shown by the diamond-shaped symbol in Figure 1.1.2(b). Current sources may also be either independent or dependent and the corresponding symbols are shown in (c) and (d). The three passive circuit elements are shown in Figure 1.1.2(e), (f), and (g).

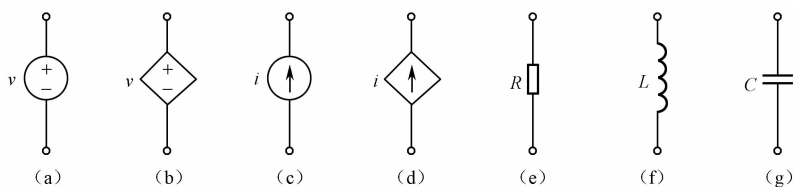


Figure 1.1.2 Seven basic circuit elements.

The circuit diagrams presented here are termed lumped-parameter circuits, since a single element in one location is used to represent a distributed resistance, inductance, or capacitance^[2]. For example, a coil consisting of a large number of turns of insulated

wire has resistance throughout the entire length of the wire. Nevertheless, a single resistance lumped at one place as in Figure 1. 1. 3(b) or (c) represents the distributed resistance. The inductance is likewise lumped at one place, either in series with the resistance as in (b) or in parallel as in (c).

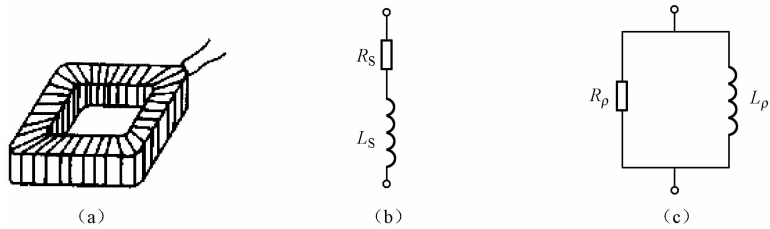


Figure 1. 1. 3 (a)A coil;(b)A single resistance;(c)The distributed resistance.

In general, a coil can be represented by either a series or a parallel arrangement of circuit elements. The frequency of the applied voltage may require that one or the other be used to represent the device.

Sign Conventions

A voltage function and a polarity must be specified to completely describe a voltage source. The polarity marks, + and −, are placed near the conductors of the symbol that identifies the voltage source. If, for example, $v = 10.0\sin\omega t$ in Figure 1. 1. 4(a), terminal A is positive with respect to B for $0 > \omega t > \pi$, and B is positive with respect to A for $\pi > \omega t > 2\pi$ for the first cycle of the sine function.

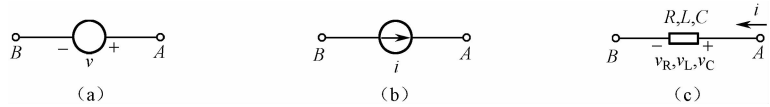


Figure 1. 1. 4 A Voltage Source,a current source and passive circuit element.

Similarly, a current source requires that a direction be indicated, as well as the function, as shown in Figure 1. 1. 4(b) [3]. For passive circuit elements R , L and C , shown in Figure 1. 1. 4(c), the terminal where the current enters is generally treated as positive with respect to the terminal where the current leaves.

The sign on power is illustrated by the DC circuit of Figure 1. 1. 5(a) with constant voltage sources $V_A = 20.0\text{ V}$ and $V_B = 5.0\text{ V}$ and a single $5\text{ }\Omega$ resistor. The resulting current of 3.0 A is in the clockwise direction. Considering now Figure 1. 1. 5(b), power is absorbed by an element when the current enters the element at the positive terminal. Power, computed by VI or I^2R , is therefore absorbed by both the resistor and the V_B source, 45.0 W and 15 W respectively. Since the current enters V_A at the negative terminal, this element is the power source for the circuit. $P = VI = 60.0\text{ W}$ confirms

that the power absorbed by the resistor and the source V_B is provided by the source V_A .

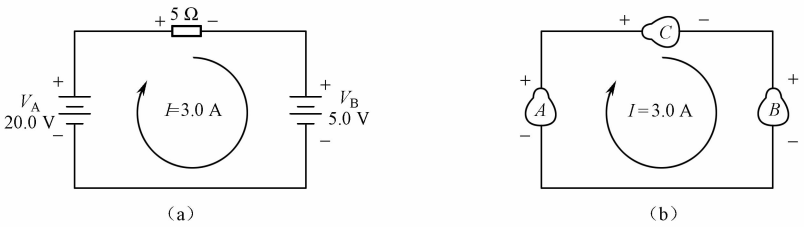
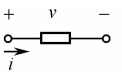
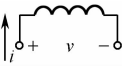
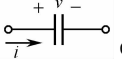


Figure 1. 1. 5 DC Circuit and its current direction.

Voltage-Current Relations

The passive circuit elements resistance R , inductance L , and capacitance C are defined by the manner in which the voltage and current are related for the individual element. For example, if the voltage v and current i for a single element are related by a constant, then the element is a resistance, R is the constant of proportionality, and $v = Ri$. Similarly, if the voltage is the time derivative of the current, then the element is an inductance, L is the constant of proportionality, and $v = Ldi/dt$. Finally, if the current in the element is the time derivative of the voltage, then the element is a capacitance, C is the constant of proportionality, and $i = Cdv/dt$. Table 1.1.1 summarizes these relationships for the three passive circuit elements. Note the current directions and the corresponding polarity of the voltages.

Table 1. 1. 1 The voltage-current relationships for the three passive circuit elements.

Circuit element	Units	Voltage	Current	Power
 Resistance	ohms (Ω)	$v = Ri$	$i = \frac{v}{R}$	$p = vi = i^2R$
 Inductance	henries(H)	$v = L \frac{di}{dt}$	$i = \frac{1}{L} \int vdt + k_1$	$p = vi = Li \frac{di}{dt}$
 Capacitance	farads (F)	$v = C \int idt + k_2$	$i = C \frac{dv}{dt}$	$p = vi = Cv \frac{dv}{dt}$

Resistance

All electrical devices that consume energy must have a resistor (also called a resistance) in their circuit model. Inductors and capacitors may store energy but over time return that energy to the source or to another circuit element. Power in the resistor, given by $p = vi = i^2R = v^2/R$, is always positive. Energy is then determined as the integral of the instantaneous power

$$w = \int_{t_1}^{t_2} p dt = R \int_{t_1}^{t_2} i^2 dt = \frac{1}{R} \int_{t_1}^{t_2} v^2 dt$$

Inductance

The circuit element that stores energy in a magnetic field is an inductor (also called an inductance). With time-variable current, the energy is generally stored during some parts of the cycle and then returned to the source during others. When the inductance is removed from the source, the magnetic field will collapse; in other words, no energy is stored without a connected source. Coils found in electric motors, transformers, and similar devices can be expected to have inductances in their circuit models. Even a set of parallel conductors exhibits inductance that must be considered at most frequencies. The power and energy relationships are as follows.

$$p = vi = L \frac{di}{dt} = dt \left[\frac{1}{2} Li^2 \right]$$
$$w_L = \int_{t_1}^{t_2} p dt = \int_{t_1}^{t_2} Li dt = \frac{1}{2} L [i_2^2 - i_1^2]$$

Energy stored in the magnetic field of an inductance is $w_L = \frac{1}{2} Li^2$.

Capacitance

The circuit element that stores energy in an electric field is a capacitor (also called capacitance). When the voltage is variable over a cycle, energy will be stored during one part of the cycle and returned in the next. While an inductance cannot retain energy after removal of the source because the magnetic field collapses, the capacitor retains the charge and the electric field can remain after the source is removed. This charged condition can remain until a discharge path is provided, at which time the energy is released. The charge, $q = Cv$, on a capacitor results in an electric field in the dielectric which is the mechanism of the energy storage^[4]. In the simple parallel-plate capacitor there is an excess of charge on one plate and a deficiency on the other. It is the equalization of these charges that takes place when the capacitor is discharged^[5]. The power and energy relationships for the capacitance are as follows.

$$p = vi = Cv \frac{dv}{dt} = \frac{d}{dt} \left[\frac{1}{2} Cv^2 \right]$$
$$w_C = \int_{t_1}^{t_2} p dt = \int_{t_1}^{t_2} Cv dv = \frac{1}{2} C [v_2^2 - v_1^2]$$

The energy stored in the electric field of capacitance is $w_C = \frac{1}{2} Cv^2$.

Circuit Diagrams

Every circuit diagram can be constructed in a variety of ways which may look different but are in fact identical. The diagram presented in a problem may not suggest the best of several methods of solution. Consequently, a diagram should be examined

before a solution is started and redrawn if necessary to show more clearly how the elements are interconnected. An extreme example is illustrated in Figure 1. 1. 6, where the three circuits are actually identical. In Figure 1. 1. 6(a) the three “junctions” labeled A are shown as two “junctions” in (b). However, resistor R_4 is bypassed by a short circuit and may be removed for purposes of analysis. Then, in Figure 1. 1. 6(c) the single junction A is shown with its three meeting branches.

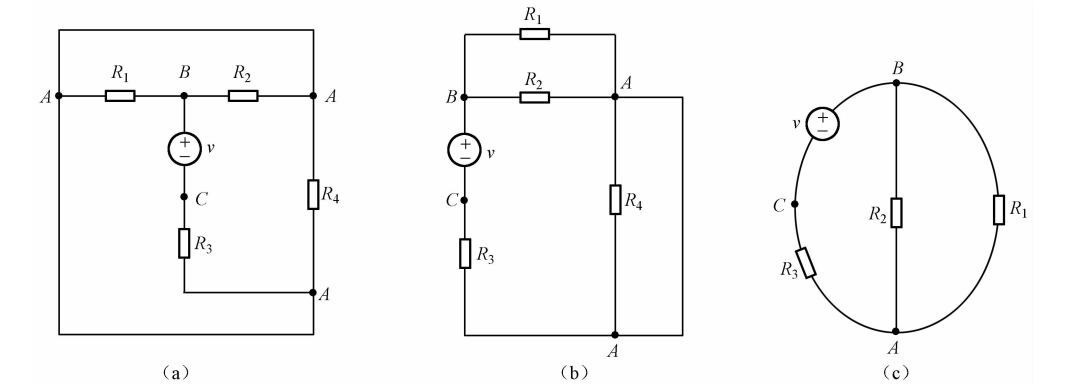


Figure 1. 1. 6 An extreme example that the three circuits are actually identical.

1. 1. 2 Specialized English Words

electrical device	电气设备	instantaneous	瞬时的
conducting lead	引线体	magnetic field	磁场
resistor	电阻器	time-variable	时变的
resistance	电阻,阻值	coil	线圈
inductor	电感器	electric motor	电动机
capacitor	电容器	transformers	变压器
capacitance	电容	dielectric	电介质
passive elements	无源元件	mechanism	机理
active elements	有源元件	paralleled-plate	平行板
illustrate	图解	charges	电荷
circuit diagram	电路图	deficiency	缺乏,不足
lumped-parameter circuits	集总参数电路	equalization	均衡,同等化
lumped	集总的	junction	节点
sign	符号	term	术语
power	功率	distributed	分布式的
sine	正弦	cycle	周期
polarity	极性	derivative	导数
energy	能量		

1. 1. 3 Notes

[1] A two-terminal element in general form is shown in Figure 1. 1. 1, with a single device represented by the rectangular symbol and two perfectly conducting leads ending at connecting points A and B. 句中“with…”为介词短语作状语修饰“is shown”。此句可译为“二端元件的一般形式如图 1. 1. 1 所示,它是用长方形符号及以 A, B 为端点的两段理想引线导体来表示的。”

[2] The circuit diagrams presented here are termed lumped-parameter circuits, since a single element in one location is used to represent a distributed resistance, inductance, or capacitance. 句中“since”后面为原因状语从句,用来说明前面的主句。此句可译为“由于我们常用位于某处单个元件来表示分布电阻、电感或电容,所以这里所列出的这些电路图称为集总参数电路。”

[3] Similarly, a current source requires that a direction be indicated, as well as the function, as shown in Figure 1. 1. 4(b). “as well as”在句中作为并列连词在科技英语中很常见,这里表示后面的“function”和前面的“direction”一样需要指明。此句可译为“与电压源类似,电流源也需要指明其方向和函数,如图 1. 1. 4(b)所示。”

[4] The charge, $q = Cv$, on a capacitor results in an electric field in the dielectric which is the mechanism of the energy storage. 句中“results in”短语表示导致,造成……结果。“which”引导的非限定性定语从句用来修饰前面的整个句子。此句可译为“电容器中的电荷, $q = Cv$, 在电介质中产生电场,这便是电容器储存能量的机理。”

[5] It is the equalization of these charges that takes place when the capacitor is discharged. 句中“It is…that”为强调句型,强调主语“the equalization of these charges”。此句可译为“当电容放电时,两极板上的电荷趋向均衡。”

1. 1. 4 Reference Translation

电路基本概念

无源元件与有源元件

电气设备是用电路图或网络来描述的,它们由串联或并联连接的两端元件组成。对电路图的 analysis 可预估实际器件的性能。二端元件的一般形式如图 1. 1. 1 所示,它是用长方形符号及以 A, B 为端点的两段理想引线体来表示的。有源元件是指能向电路网络提供能量的电压源或电流源。电阻、电感和电容是无源元件,它们吸收电源能量,并将这些能量转化为其他形式或将它存储在电场或磁场中。

图 1. 1. 2 列举了七种基本电路元件。元件(a)和(b)是电压源,(c)和(d)是电流源。图 1. 1. 2(a)所示圆圈表示的是独立电压源,它不受所连接电路变化的影响。受控电压源随所连接电路中描述的条件按一定的方式变化,以图 1. 1. 2(b)所示的菱形表示。电流源也分为独立电流源和受控电流源,相应的符号分别示于(c)和(d)。图 1. 1. 2(e)、(f)和(g)所示为三种无源电路元件的符号。

由于我们常用位于某处的单个元件来表示分布电阻、电感或电容,所以这里所列出的这些电路图称为集总参数电路。比如,由绝缘导线绕制的多匝线圈在整个导线中都分布有电阻,而电路图 1.1.3(b)和(c)中用集中在一个地方的电阻来表示在电路中的分布电阻,而电感也做同样的集中处理,在电路中与电阻串联[如图(b)所示]或并联[如图(c)所示]。

一般来说,线圈可以用电路元件的串联或并联形式表示。根据所加电压的频率需要用这两种形式之一表示器件。

符号约定

为完整地描述电压源,必须指出电压函数与极性。极性用符号+,−表示,置于电压源符号中导线的旁边。例如图 1.1.4(a)中, $v = 10.0\sin\omega t$, 在正弦函数的第一个周期,当 $0 > \omega t > \pi$ 时,端点 A 相对 B 为正,而当 $\pi > \omega t > 2\pi$ 时,端点 B 相对 A 为正。

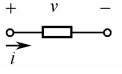
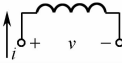
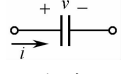
与电压源类似,电流源也需要指明其方向和函数,如图 1.1.4(b)所示。在图 1.1.4(c)所示的无源元件 R 、 L 和 C 中,电流流入的一端设为正极性端,而电流流出的一端则为负极性端。

用图 1.1.5(a)的直流电路来说明电源的符号,电路由直流电压源 $V_A = 20.0\text{ V}$, $V_B = 5.0\text{ V}$ 和一个 $5\text{ }\Omega$ 电阻组成。电路中所得的电流为 3.0 A , 呈顺时针方向。就图 1.1.5(b)来说,当电流从正极性端点流入时,元件吸收功率。可用 VI 或 I^2R 来计算功率,这两部分功率被电压源 V_B 和电阻吸收,分别为 45.0 W 和 15 W 。由于电流从 V_A 的负端流入,这一元件为电路提供功率, $P = VI = 60.0\text{ W}$ 表明由电阻和电源 V_B 吸收的功率是由电源 V_A 提供的。

电压-电流关系

无源电路元件电阻 R 、电感 L 和电容 C , 是以各元件中电压和电流相关联的形式进行定义的。例如,若单个元件的电压 v 与电流 i 的关系为一常量,则此元件为电阻, R 为常比例系数,且 $v = Ri$ 。类似地,若电压是电流对时间的导数, $v = Ldi/dt$, L 为常比例系数,则此元件为电感。最后,若元件中的电流是电压对时间的导数, $i = Cdv/dt$, C 为常比例系数,则该元件为电容。表 1.1.1 归纳了这三种无源元件的电压-电流关系。注意电流的方向和相应电压的极性。

表 1.1.1 三种无源电路元件的电压-电流关系

电 路 元 件	单 位	电 压	电 流	功 率
 电 阻	ohms (Ω)	$v = Ri$	$i = \frac{v}{R}$	$p = vi = i^2R$
 电 感	henries (H)	$v = L \frac{di}{dt}$	$i = \frac{1}{L} \int v dt + k_1$	$p = vi = Li \frac{di}{dt}$
 电 容	farads (F)	$v = C \int i dt + k_2$	$i = C \frac{dv}{dt}$	$p = vi = Cv \frac{dv}{dt}$

电阻元件

所有耗能的器件在它的电路模型中必有电阻。电感器和电容器可以存储能量,并随后将能量返回给电源或其他元件。电阻器中的功率总为正,可用公式 $p = vi = i^2 R = v^2 R$ 计算。电能可由瞬时功率的积分来确定:

$$w = \int_{t_1}^{t_2} P dt = R \int_{t_1}^{t_2} i^2 dt = \frac{1}{R} \int_{t_1}^{t_2} v^2 dt$$

电感元件

在磁场中存储能量的电路元件称为电感器(也称为电感)。在时变电流的作用下,电感在一个周期的一段时间里存储能量,而在其他时间段里又释放能量给电源。当电感元件脱离电源时,磁场将不存在,即电感元件如果不连接电源就不能存储能量。在电动机、变压器及类似装置中,有线圈的电路模型都有电感。在大多数频率下,甚至对一组平行导体呈现出的电感也必须予以考虑。电感的功率和能量的关系如下所示:

$$p = vi = L \frac{di}{dt} i = \frac{d}{dt} \left[\frac{1}{2} Li^2 \right]$$
$$w_L = \int_{t_1}^{t_2} p dt = \int_{t_1}^{t_2} Li di = \frac{1}{2} L [i_2^2 - i_1^2]$$

电感元件在磁场中的储能为 $w_L = \frac{1}{2} Li^2$ 。

电容元件

在电场中存储能量的电路元件称为电容器(也称为电容)。当电压在一个周期中变化时,在周期的一段时间内会存储能量,而在随后的时间内又将其释放出来。当去掉电源时,电感元件由于磁场消失而不能保存能量,而电容元件能存储电荷,所以在去掉电源后,电场仍能保持。这种充电的状态将一直保持,直到有放电回路时能量才会释放。电容器中的电荷, $q = Cv$, 在电介质中产生电场,这便是电容器存储能量的机理。在简单平板电容器中,一个极板电荷过剩,而另一个极板电荷不足。当电容放电时,两极板上的电荷趋向均衡。电容元件的功率与能量的关系如下所示:

$$p = vi = Cv \frac{dv}{dt} = \frac{d}{dt} \left[\frac{1}{2} Cv^2 \right]$$
$$w_C = \int_{t_1}^{t_2} p dt = \int_{t_1}^{t_2} Cv dv = \frac{1}{2} C [v_2^2 - v_1^2]$$

电容元件在电场中的储能为 $w_C = \frac{1}{2} Cv^2$ 。

电路图

每一电路的电路图可以有多种形式,它们看上去虽然各不相同,但其实质都一样。表示一个问题的电路图未必是求解问题的最佳形式。因此,解决问题前应首先研究电路图,如有必要可重画电路图,以便更清楚地表示元件的连接关系。图 1.1.6 给出了一个很明显的例子,其中三个电路实际上是等效的。在图 1.1.6(a)中,标为 A 的三个“连接点”在

(b)中为两个节点。而在进行电路分析时,电阻 R_4 由于被短路线旁路,可去掉。于是就在图 1.1.6(c)中画出单节点 A 为三条支路的交点。

1.1.5 Reading Materials

Nonlinear (非线性) Resistors

The current-voltage relationship in an element may be instantaneous but not necessarily linear. The element is then modeled as a nonlinear resistor. An example is a filament lamp (白炽灯) which at higher voltages draws proportionally less current. Another important electrical device modeled as a nonlinear resistor is a diode. A diode is a two-terminal device that, roughly speaking, conducts electric current in one direction [from anode to cathode, called forward-biased (正向偏置)] much better than the opposite direction (reverse-biased). The circuit symbol for the diode and an example of its current-voltage characteristic are shown in Figure 1.1.7. The arrow is from the anode to the cathode and indicates the forward direction ($i > 0$). A small positive voltage at the diode's terminal biases the diode in the forward direction and can produce a large current. A negative voltage biases the diode in the reverse direction and produces little current even at large voltage values. An ideal diode is a circuit model which works like a perfect switch. See Figure 1.1.8. Its (i, v) characteristic is

$$\begin{cases} v = 0 & \text{when } i \geqslant 0 \\ i = 0 & \text{when } v \leqslant 0 \end{cases}$$

The static resistance of a nonlinear resistor operating at (I, V) , is $R = V/I$. Its dynamic resistance is $r = \Delta V/\Delta I$ which is the inverse of the slope(斜率) of the current plotted versus voltage. Static and dynamic resistances both depend on the operating

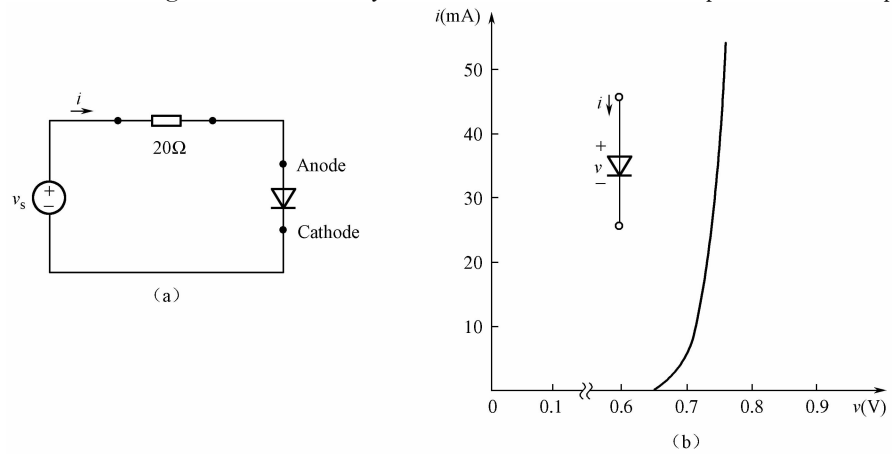


Figure 1.1.7 Diode's current-voltage characteristic.

point.

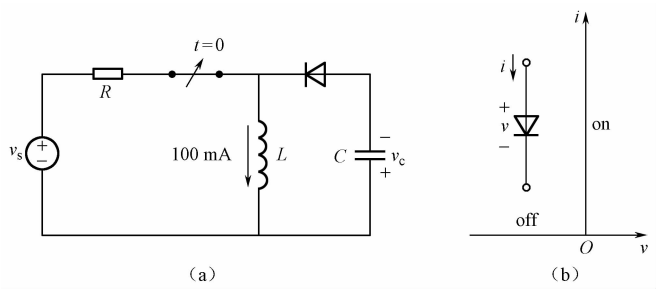


Figure 1. 1. 8 Circuit model and its(I,V) characteristic

1. 2 Voltage and Current Laws

1. 2. 1 Text

Analysis of linear circuits rests on two fundamental physical laws that describe how the voltages and currents in a circuit must behave. This behavior results from whatever voltage sources, current sources, and energy storage elements are connected to the circuit^[1]. A voltage source imposes a constraint on the evolution of the voltage between a pair of nodes; a current source imposes a constraint on the evolution of the current in a branch of the circuit. The energy storage elements (capacitors and inductors) impose initial conditions on currents and voltages in the circuit; they also establish a dynamic relationship between the voltage and the current at their terminals.

Regardless of how a linear circuit is stimulated, every node voltage and every branch current, at every instant of time, must be consistent with Kirchhoff’s voltage and current laws. These two laws govern even the most complex linear circuits. (They also apply to a broad category of nonlinear circuits that are modeled by point models of voltage and current.)

A circuit can be considered to have a topological (or graph) view, consisting of a labeled set of nodes and a labeled set of edges. Each edge is associated with a pair of nodes. A node is drawn as a dot and represents a connection between two or more physical components; an edge is drawn as a line and represents a path, or branch, for current flow through a component (see Figure 1. 2. 1).

The edges, or branches, of the graph are assigned current labels, i_1, i_2, \dots, i_m . Each current has a designated direction, usually denoted by an arrow symbol. If the arrow is drawn toward a node, the associated current is said to be entering the node; if the arrow is drawn away from the node, the current is said to be leaving the node. The current i_1 is entering node b in Figure 1. 2. 1; the current i_5 is leaving node e .

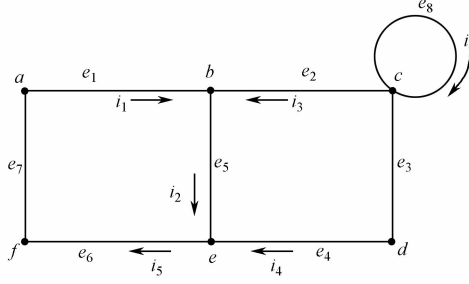


Figure 1.2.1 Graph representation of a linear circuit.

Given a branch, the pair of nodes to which the branch is attached defines the convention for measuring voltages in the circuit^[2]. Given the ordered pair of nodes (a, b) , a voltage measurement is formed as follows:

$$v_{ab} = v_a - v_b$$

where v_a and v_b are the absolute electrical potentials (voltages) at the respective nodes, taken relative to some reference node^[3]. Typically, one node of the circuit is labeled as ground, or reference node; the remaining nodes are assigned voltage labels. The measured quantity, v_{ab} , is called the voltage drop from node a to node b . We note that

$$v_{ab} = -v_{ba}$$

and that

$$v_{ba} = v_b - v_a$$

is called the voltage rise from a to b . Each node voltage implicitly defines the voltage drop between the respective node and the ground node.

The pair of nodes to which an edge is attached may be written as (a, b) or (b, a) . Given an ordered pair of nodes (a, b) , a path from a to b is a directed sequence of edges in which the first edge in the sequence contains node label a , the last edge in the sequence contains node label b , and the node indices of any two adjacent members of the sequence have at least one node label in common. In Figure 1.2.1, the edge sequence $\{e_1, e_2, e_4\}$ is not a path, because e_2 and e_4 do not share a common node label. The sequence $\{e_1, e_2\}$ is a path from node a to node c .

A path is said to be closed if the first node index of its first edge is identical to the second node index of its last edge. The following edge sequence forms a closed path in the graph given in Figure 1.2.1: $\{e_1, e_2, e_3, e_4, e_6, e_7\}$. Note that the edge sequences $\{e_8\}$ and $\{e_1, e_1\}$ are closed paths.

Kirchhoff's Current Law

Kirchhoff's current law (KCL) imposes constraints on the currents in the branches that are attached to each node of a circuit. In simplest terms, KCL states that the sum of the currents that are entering a given node must equal the sum of the currents that are

leaving the node. Thus, the set of currents in branches attached to a given node can be partitioned into two groups whose orientation is away from (into) the node. The two groups must contain the same net current. Applying KCL at node b in Figure 1.2.1 gives

$$i_1(t) + i_3(t) = i_2(t)$$

A connection of water pipes that has no leaks is a physical analogy of this situation. The net rate at which water is flowing into a joint of two or more pipes must equal the net rate at which water is flowing away from the joint. The joint itself has the property that it only connects the pipes and thereby imposes a structure on the flow of water, but it cannot store water. This is true regardless of when the flow is measured. Likewise, the nodes of a circuit are modeled as though they cannot store charge. (Physical circuits are sometimes modeled for the purpose of simulation as though they store charge, but these nodes implicitly have a capacitor that provides the physical mechanism for storing the charge. Thus, KCL is ultimately satisfied.)

KCL can be stated alternatively as: “the algebraic sum of the branch currents entering (or leaving) any node of a circuit at any instant of time must be zero.” In this form, the label of any current whose orientation is away from the node is preceded by a minus sign. The currents entering node b in Figure 1.2.1 must satisfy

$$i_1(t) - i_2(t) + i_3(t) = 0$$

In general, the currents entering or leaving each node m of a circuit must satisfy

$$\sum i_{km}(t) = 0$$

where $i_{km}(t)$ is understood to be the current in branch k attached to node m . The currents used in this expression are understood to be the currents that would be measured in the branches attached to the node, and their values include a magnitude and an algebraic sign^[4]. If the measurement convention is oriented for the case where currents are entering the node, then the actual current in a branch has a positive or negative sign, depending on whether the current is truly flowing toward the node in question.

Once KCL has been written for the nodes of a circuit, the equations can be rewritten by substituting into the equations the voltage-current relationships of the individual components^[5]. If a circuit is resistive, the resulting equations will be algebraic. If capacitors or inductors are included in the circuit, the substitution will produce a differential equation. For example, writing KCL at the node for v_3 in Figure 1.2.2 produces

$$i_2 + i_1 - i_3 = 0$$

and

$$C_1 \frac{dv_1}{dt} + \frac{v_4 - v_3}{R_2} - C_2 \frac{dv_2}{dt} = 0$$

KCL for the node between C_2 and R_1 can be written to eliminate variables and lead to a solution describing the capacitor voltages. The capacitor voltages, together with the applied voltage source, determine the remaining voltages and currents in the circuit. Nodal analysis treats the systematic modeling and analysis of a circuit under the influence of its sources and energy storage elements.

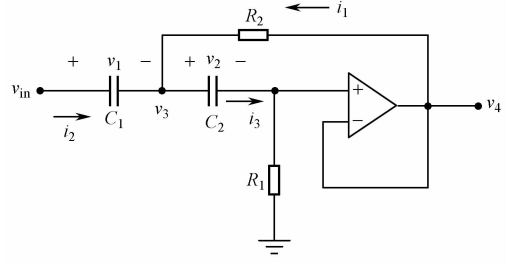


Figure 1. 2. 2 Example of a circuit containing energy storage elements.

Kirchhoff's Voltage Law

Kirchhoff's voltage law (KVL) describes a relationship among the voltages measured across the branches in any closed, connected path in a circuit. Each branch in a circuit is connected to two nodes. For the purpose of applying KVL, a path has an orientation in the sense that in "walking" along the path one would enter one of the nodes and exit the other. This establishes a direction for determining the voltage across a branch in the path; the voltage is the difference between the potential of the node entered and the potential of the node at which the path exits. Alternatively, the voltage drop along a branch is the difference of the node voltage at the entered node and the node voltage at the exit node. For example, if a path includes a branch between node a and node b , the voltage drop measured along the path in the direction from node a to node b is denoted by v_{ab} and is given by $v_{ab} = v_a - v_b$ ^[6]. Given v_{ab} , branch voltage along the path in the direction from node b to node a is $v_{ba} = v_b - v_a$.

Kirchhoff's voltage law, like Kirchhoff's current law, is true at any time. KVL can also be stated in terms of voltage rises instead of voltage drops.

KVL can be expressed mathematically as "the algebraic sum of the voltages drops around any closed path of a circuit at any instant of time is zero." This statement can also be cast as an equation:

$$\sum v_{km}(t) = 0$$

where $v_{km}(t)$ is the instantaneous voltage drop measured across branch k of path m . By convention, the voltage drop is taken in the direction of the edge sequence that forms the path.

The edge sequence $\{e_1, e_2, e_3, e_4, e_6, e_7\}$ forms a closed path in Figure 1. 2. 1. The sum of the voltage drops taken around the path must satisfy KVL:

$$v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t) + v_{fa}(t) = 0$$

Since $v_{af} = -v_{fa}$, we can also write

$$v_{af}(t) = v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t)$$

Had we chosen the path corresponding to the edge sequence $\{e_1, e_5, e_6, e_7\}$ for the path, we would have obtained

$$v_{af}(t) = v_{ab}(t) + v_{bc}(t) + v_{ef}(t)$$

This demonstrates how KCL can be used to determine the voltage between a pair of nodes. It also reveals the fact that the voltage between a pair of nodes is independent of the path between the nodes on which the voltages are measured.

1. 2. 2 Specialized English Words

linear circuit	线性电路	variable	变量
law	定律	potential	电位
stimulate	激励	voltage drop	电压降
node	节点	Kirchhoff's Current Law	基尔霍夫电流定律
order	顺序	Kirchhoff's Voltage Law	基尔霍夫电压定律
branch	支路	differential equation	微分方程

1. 2. 3 Notes

[1] This behavior results from whatever voltage sources, current sources, and energy storage elements are connected to the circuit. 句中“whatever”在这里引导名词从句,表示任何电压源、电流源和储能元件。短语“result from”表示由……产生”。全句翻译为“这种规律与电压源、电流源和储能元件具体的类型无关。”

[2] Given a branch, the pair of nodes to which the branch is attached defines the convention for measuring voltages in the circuit. 句中“to which the branch is attached”为定语从句,修饰“the pair of nodes”。全句翻译为“对于电路中给定支路两端电压的测量,是由支路两端的节点来确定的。”

[3] Given the ordered pair of nodes (a, b) , a voltage measurement is formed as follows: $v_{ab} = v_a - v_b$, where v_a and v_b are the absolute electrical potentials(voltage) at the respective nodes, taken relative to some reference node. 句中的“taken relative to some reference node”是过去分词短语,用来修饰“the absolute electrical potentials”。全句翻译为“对于给定顺序的一对节点 (a, b) ,测量其电压量应为 $v_{ab} = v_a - v_b$, 式中 v_a 和 v_b 表示相对于某参考节点电势的绝对值(电压)。”

[4] The currents used in this expression are understood to be the currents that would be measured in the branches attached to the node, and their values include a magnitude and an algebraic sign. 句中“to be the currents”为不定式短语做结果状语,而“that would be…”为定语从句,用来修饰“current”;“attached to the node”为过去分词短语,用来修饰“the branches”。全句翻译为“式中的电流为连接到节点的支路电流,它们的

值包括大小和符号。”

[5]Once KCL has been written for the nodes of a circuit, the equations can be rewritten by substituting into the equations the voltage-current relationships of the individual components. 句中“the voltage-current relationships …”是“substituting”的宾语,等于“by substituting i the voltage-current relationships of the individual components into the equations”。全句翻译为“当写好电路节点的 KCL 方程后,我们可以把每个元件的电压电流关系代入节点的 KCL 方程,而得到新的公式。”

[6]For example, if a path includes a branch between node a and node b , the voltage drop measured along the path in the direction from node a and node b is denoted by v_{ab} and is given by $v_{ab} = v_a - v_b$. 句中“measured along the path in the direction from node a and node b ”是过去分词短语做后置定语,用来说明“the voltage drop”。全句翻译为“例如,如果某通路包括含节点 a 和节点 b 的支路,电压降的方向是从 a 到 b ,我们标识为 v_{ab} , $v_{ab} = v_a - v_b$ 。”

1. 2. 4 Reference Translation

电压与电流定律

描述电路中电压和电流规律的两个基本物理定律是分析线性电路的基础。这种规律适用于电路中任何电压源、电流源和储能元件。这种规律与电压源、电流源和储能元件具体的类型无关。加在一对节点两端的电压决定了其电压变化规律;支路中的电流源决定了该支路的电流变化规律。储能元件(电容和电感)决定了电路中的电压和电流的初始条件;同时确定了元件两端电压和电流之间的动态关系。

无论线性电路是如何被激励的,每个节点电压和每个支路电流,在任何时候都必须遵循基尔霍夫电压和电流定律。这两条定律适用于任何复杂的线性电路。(也适用于广义的非线性电路,其中电压和电流用点的模型来描述。)

我们可以用一组带标记的节点和一组带标记的边的拓扑图的形式来描述电路。每个边连接一对节点。节点用点来表示,代表两个或者多个物理元件之间的连接。边用线来表示,代表电流流过某个物理元件的一个通路或者一个支路。

图中边或者支路用电流符号 i_1, i_2, \dots, i_m 来标识。每个电流都有指定的方向,通常用箭头符号来表示。如果箭头指向节点,那么我们说对应的那个电流是流入节点;如果箭头背向节点,那么对应的电流则是流出节点。在图 1. 2. 1 中,电流 i_1 流入节点 b , 电流 i_5 流出节点 e 。

对于电路中给定支路两端电压的测量,是由支路两端的节点来确定的。对于给定顺序的一对节点(a, b),测量其电压量应为

$$v_{ab} = v_a - v_b$$

式中 v_a 和 v_b 表示相对于某参考节点电势的绝对值(电压)。通常电路中有一个节点为接地,或者参考节点;其余的节点用电压来标识。我们称电压 v_{ab} 的大小为节点 a 到节点 b

的电压降。这里 $v_{ab} = -v_{ba}$ ，而 $v_{ba} = v_b - v_a$ 称为节点 a 到节点 b 的电压升。每个节点电压隐含表示了它对地的电压降。

我们可以把分别接有边的节点对写为 (a, b) 或者 (b, a) 。对于有先后顺序的节点 (a, b) ，从 a 到 b 的通路是一个有向的边的序列，其中包含节点 a 的边为第一个边，最后一个边包含节点 b ；边序列中任意两个相邻边至少有一个公共节点。在图 1.2.1 中，由于 e_2 和 e_4 没有公共节点，所以边序列 $\{e_1, e_2, e_3\}$ 不构成一个通路。而边序列 $\{e_1, e_2\}$ 构成节点 a 到节点 c 一个通路。

如果节点序列中的第一个边的第一个节点与最后一个边的第二个节点是同一个节点，那么我们说这个通路是封闭的。在图 3.1 中的边 $\{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ 构成一个封闭通路。这里我们注意到，边序列 $\{e_8\}$ 和 $\{e_1, e_1\}$ 也是封闭通路。

基尔霍夫电流定律

基尔霍夫电流定律(KCL)规定了连接到电路的每个节点的分支电流的规律。简单地说，KCL 叙述的是，流入一个已知节点的电流的总和与流出该节点的电流总和必须相等。这样，连接到一个已知节点的支路电流可以分为两组，其方向分别为流入节点和流出节点。两组电流包含的必须是相同的净电流。对图 1.2.1 中的节点 b 应用 KCL 定律有

$$i_1(t) + i_3(t) = i_2(t)$$

无缝的水管连接与我们这里的情况在物理上非常类似。流入两个或者多个水管的连接点的净速率和流出该连接点的水的净速率必须相等。节点本身的特性就是它仅仅连接水管，为水的流动建立了一个结构而不存储水，与有没有水流无关。同样，电路的节点也不存储电荷，它们也可以同样地建立相应的结构。（电路物理模型有时为了仿真的需要认为节点存储电荷，不过这些节点往往含有电容，它为存储电荷提供了物理机制。因此，KCL 定律最终还是可以满足的。）

换一种说法，我们还可以将基尔霍夫电流定律描述为：“在任何瞬间，对于电路中的任意一个节点，流入（或流出）电路中任何一个节点的电流的代数和为零。”按照这种描述方法，我们把流出节点的电流标记为负号。在图 1.2.1 中流入节点 b 的电流必须满足公式

$$i_1(t) - i_2(t) + i_3(t) = 0$$

总的来说，对于电路中的节点 m ，流入或者流出的电流必须满足

$$\sum i_{km}(t) = 0$$

该式中 $i_{km}(t)$ 表示连接到节点 m 的支路 k 的电流。式中的电流为连接到节点的支路电流，它们的值包括大小和符号。如果测量的方法是根据节点电流如何流入节点来确定的，那么实际支路中的电流就有正负之分，具体取决于电流是否为流向节点。

当写好电路节点的 KCL 方程后，我们可以把每个元件的电压电流关系代入节点的 KCL 方程，而得到新的公式。如果电路为电阻性的，得到的方程的将为代数方程。如果电路中有电容或者电感元件，那么将会得到微分方程。例如，在图 1.2.2 中，我们对节点 v_3 写基尔霍夫电流方程，则有

$$i_2 + i_1 - i_3 = 0$$

与

$$C_1 \frac{dv_1}{dt} + \frac{v_4 - v_3}{R_2} - C_2 \frac{dv_2}{dt} = 0$$

列出 C_2 和 R_1 之间的节点的 KCL 方程,可以消除变量进而求解得出电容上的电压。由电容上的电压和施加的电压源可以确定电路中其他电压和电流。将节点分析法用于系统的建模和电路分析,会受到系统的电源和储能元件的影响。

基尔霍夫电压定律

基尔霍夫电压定律(KVL)描述的是在电路中的任何一个闭合回路中的支路电压之间的关系。电路中每个支路都有 2 个节点。为了应用 KVL 定律,我们按此种意义规定一种支路的绕行方向,即从一个节点进入,另一个节点离开。这样就确立了支路中的电压的方向,该电压为进入节点和离开节点的电势差,或者说沿支路方向的电压降即为进入节点和离开节点的电压差。例如,如果某通路包括含节点 a 和节点 b 的支路,电压降的方向是从 a 到 b ,我们标识为 v_{ab} , $v_{ab} = v_a - v_b$ 。相对于 v_{ab} ,在支路中从 b 到 a 的支路电压为 $v_{ba} = v_b - v_a$ 。

同基尔霍夫电流定律一样,基尔霍夫电压定律在任何时候都是成立的。KVL 定律也可以不用电压降,而用电压升来描述。

可用数学的方法将 KVL 定律描述为:“电路中沿任意闭合回路的电压降在任何时候的代数和为零”。我们同样可以用公式来表达

$$\sum v_{km}(t) = 0$$

式中 $v_{km}(t)$ 为回路 m 中支路 k 的瞬时电压降。通常我们取构成通路的边序列的方向为电压降的方向。

图 1.2.1 中,边序列 $\{e_1, e_2, e_3, e_4, e_6, e_7\}$ 构成一个闭合回路。沿回路电压降的和必须满足 KVL 定律

$$v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t) + v_{fa}(t) = 0$$

由于 $v_{af} = -v_{fa}$,我们也可以写为

$$v_{af}(t) = v_{ab}(t) + v_{bc}(t) + v_{cd}(t) + v_{de}(t) + v_{ef}(t)$$

如果我们选择边序列 $\{e_1, e_5, e_6, e_7\}$ 作为回路,那么可以得到

$$v_{af}(t) = v_{ab}(t) + v_{bc}(t) + v_{ef}(t)$$

这表明 KVL 可以用来确定一对节点之间的电压。同时也说明两节点之间的电压与回路无关。

1.2.5 Reading Materials

Importance of KVL and KCL

Kirchhoff's current law is used extensively in nodal analysis(节点分析法) because it is amenable to computer-based implementation and supports a systematic approach to

circuit analysis. Nodal analysis leads to a set of algebraic equations in which the variables are the voltages at the nodes of the circuit. This formulation is popular in CAD programs because the variables correspond directly to physical quantities that can be measured easily.

Kirchhoff’s voltage law can be used to completely analyze a circuit, but it is seldom used in large-scale circuit simulation programs. The basic reason is that the currents that correspond to a loop of a circuit do not necessarily correspond to the currents in the individual branches of the circuit. Nonetheless, KVL is frequently used to troubleshoot a circuit by measuring voltage drops across selected components.

1.3 Network Theorems

1.3.1 Text

Linearity

Consider a system (which may consist of a single network element) represented by a block, as shown in Figure 1.3.1, and observe that the system has an input designated by e (for excitation) and an output designated by r (for response). The system is considered to be linear if it satisfies the homogeneity and superposition conditions.

The homogeneity condition: If an arbitrary input to the system, e , causes a response, r , then if ce is the input, the output is cr , where c is some arbitrary constant.

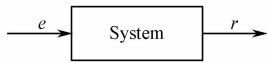


Figure 1.3.1 A simple system.

The superposition condition: If the input to the system, e_1 , causes a response, r_1 , and if an input to the system, e_2 , causes a response, r_2 , then a response, $r_1 + r_2$, will occur when the input

is $e_1 + e_2$.

If neither the homogeneity condition nor the superposition condition is satisfied, the system is said to be nonlinear.

The Superposition Theorem

While both the homogeneity and superposition conditions are necessary for linearity, the superposition, in itself, provides the basis for the superposition theorem:

If cause and effect are linearly related, the total effect due to several causes acting simultaneously is equal to the sum of the individual effects due to each of the causes acting one at a time^[1].

Example 1.3.1

Consider the network driven by a current source at the left and a voltage source at the top, as shown in Figure 1.3.2(a). The current phasor indicated by \hat{I} is to be

determined. According to the superposition theorem, the current \hat{I} will be the sum of the two current components \hat{I}_V due to the voltage source acting alone as shown in Figure 1. 3. 2(b) and \hat{I}_C due to the current source acting alone shown in Figure 1. 3. 2(c) .

$$\hat{I} = \hat{I}_V + \hat{I}_C$$

Figure 1. 3. 2 (b) and (c) follow from the methods of removing the effects of independent voltage and current sources. Voltage sources are nulled in a network by replacing them with short circuits and current sources are nulled in a network by replacing them with open circuits.

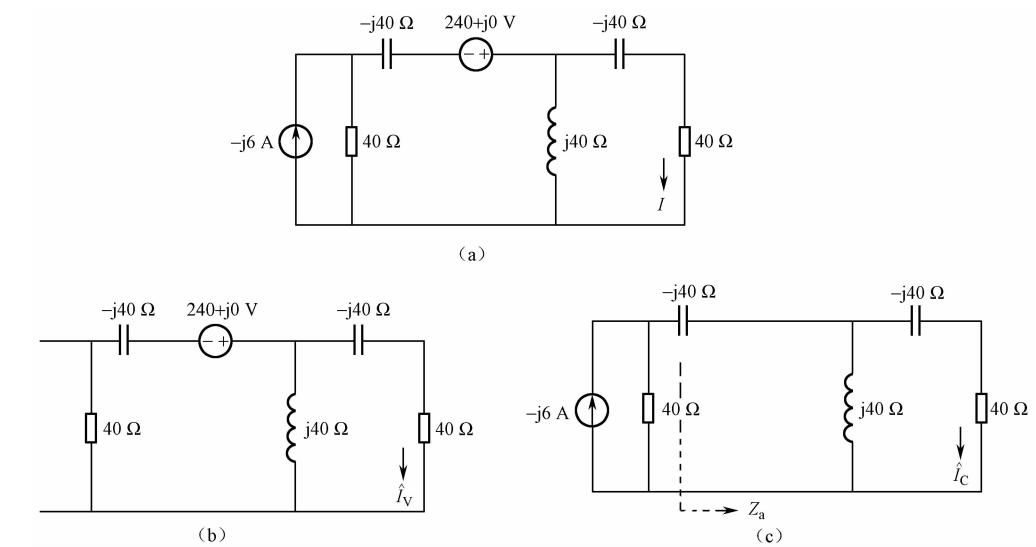


Figure 1. 3. 2 (a) A network to be solved by using superposition; (b) the network with the current source nulled; and (c) the network with the voltage source nulled.

The networks displayed in Figure 1. 3. 2(b) and (c) are simple ladder networks in the phasor domain, and the strategy is to first determine the equivalent impedances presented to the voltage and current sources. In Figure 1. 3. 2(b), the group of three impedances to the right of the voltage source are in series-parallel and possess an impedance of

$$Z_P = \frac{(40 - j40)(40j)}{40 + 40j - j40} = 40 + j40\ \Omega$$

and the total impedance presented to the voltage source is

$$Z = Z_P + 40 - j40 = 40 + j40 + 40 - j40 = 80\ \Omega$$

Then \hat{I}_1 , the current leaving the voltage source, is

$$\hat{I}_1 = \frac{240 + j0}{80} = 3 + 0j\text{ A}$$

and by a current division

$$\hat{I}_V = \left[\frac{j40}{40 - j40 + j40} \right] (3 + 0j) = j(3 + j0) = 0 + 3j \text{ A}$$

In Figure 1.3.2(b), the current source delivers current to the 40Ω resistor and to an impedance consisting of the capacitor and Z_P . Call this impedance Z_a so that

$$Z_a = -j40 + Z_P = -j40 + j40 + 40 + j40 = 40 \Omega$$

Then, two current divisions give \hat{I}_C

$$\hat{I}_C = \left[\frac{40}{40 + 40} \right] \left[\frac{j40}{40 - j40 + j40} \right] (0 + 6j) = \frac{j}{2} (0 + 6j) = 3 + 0j \text{ A}$$

The current I in the circuit of Figure 1.3.2(a) is

$$\hat{I} = \hat{I}_V + \hat{I}_C = 0 + j3 + (3 + j0) = 3 + 3j \text{ A}$$

The Network Theorems of Thévenin and Norton

If interest is to be focused on the voltages and across the currents through a small portion of a network such as network B in Figure 1.3.3(a), it is convenient to replace network A, which is complicated and of little interest, by a simple equivalent. The simple equivalent may contain a single, equivalent, voltage source in series with an equivalent impedance in series as displayed in Figure 1.3.3(b). In this case, the equivalent is called a Thévenin equivalent. Alternatively, the simple equivalent may consist of an equivalent current source in parallel with an equivalent impedance. This equivalent, shown in Figure 1.3.3(c), is called a Norton equivalent. Observe that as long as Z_T (subscript T for Thévenin) is equal to Z_N (subscript N for Norton), the two equivalents may be obtained from one another by a simple source transformation.

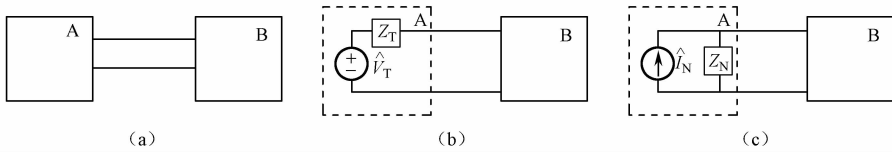


Figure 1.3.3 (a) Two one-port networks; (b) the Thévenin equivalent for network A; and (c) the Norton equivalent for network A.

Conditions of Application

The Thévenin and Norton network equivalents are only valid at the terminals of network A in Figure 1.3.3(a) and they do not extend to its interior. In addition, there are certain restrictions on networks A and B. Network A may contain only linear elements but may contain both independent and dependent sources. Network B, on the other hand, is not restricted to linear elements; it may contain nonlinear or time-varying elements and may also contain both independent and dependent sources. Together, there can be no controlled source coupling or magnetic coupling between networks A and B.

The Thévenin Theorem

The statement of the Thévenin theorem is based on Figure 1.3.3(b).

Insofar as a load which has no magnetic or controlled source coupling to a one-port is concerned, a network containing linear elements and both independent and controlled sources may be replaced by an ideal voltage source of strength, \hat{V}_T , and an equivalent impedance Z_T , in series with the source^[2]. The value of \hat{V}_T is the open-circuit voltage, \hat{V}_{oc} , appearing across the terminals of the network and Z_T is the driving point impedance at the terminals of the network, obtained with all independent sources set equal to zero.

The Norton Theorem

The Norton theorem involves a current source equivalent. The statement of the Norton theorem is based on Figure 1.3.3(c).

Insofar as a load which has no magnetic or controlled source coupling to a one-port is concerned, the network containing linear elements and both independent and controlled sources may be replaced by an ideal current source of strength, \hat{I}_N , and an equivalent impedance, Z_N , in parallel with the source. The value of \hat{I}_N is the short-circuit current, \hat{I}_{sc} , which results when the terminals of the network are shorted and Z_N is the driving point impedance at the terminals when all independent sources are set equal to zero.

The Equivalent Impedance, $Z_T = Z_N$

Three methods are available for the determination of Z_T . All of them are applicable at the analyst's discretion. When controlled sources are present, however, the first method cannot be used.

The first method involves the direct calculation of $Z_{eq} = Z_T = Z_N$ by looking into the terminals of the network after all independent sources have been nulled. Independent sources are nulled in a network by replacing all independent voltage sources with a short circuit and all independent current sources with an open circuit.

The second method, which may be used when controlled sources are present in the network, requires the computation of both the Thévenin equivalent voltage (the open-circuit voltage at the terminals of the network) and the Norton equivalent current (the current through the short-circuited terminals of the network). The equivalent impedance is the ratio of these two quantities

$$Z_T = Z_N = Z_{eq} = \frac{\hat{V}_T}{\hat{I}_N} = \frac{\hat{V}_{oc}}{\hat{I}_{sc}}$$

The third method may also be used when controlled sources are present within the network. A test voltage may be placed across the terminals with a resulting current

calculated or measured. Alternatively, a test current may be injected into the terminals with a resulting voltage determined. In either case, the equivalent resistance can be obtained from the value of the ratio of the test voltage \hat{V}_o to the resulting current \hat{I}_o ,

$$Z_T = \frac{\hat{V}_o}{\hat{I}_o}$$

1.3.2 Specialized English Words

linearity	线性	simultaneously	同时的
excitation	激励	short circuit	短路
response	响应	open circuit	开路
homogeneity	齐次性	driving point impedance	策动点阻抗
superposition	可加性	thévenin Theorem	戴维南定理
arbitrary	任意的	Norton Theorem	诺顿定理
constant	常数	equivalent impedance	等效阻抗
nonlinear	非线性		

1.3.3 Notes

[1]If cause and effect are linearly related, the total effect due to several causes acting simultaneously is equal to the sum of the individual effects due to each of the causes acting one at a time. 句中“cause and effect”本意为“因果”，这里指“激励和响应”。短语“due to”表示“归因于……”。“at a time”表示“每次”。此句可译为“假设激励和响应是线性相关的,那么多个激励同时作用在系统上产生的总的响应等于每个激励分别单独作用时产生的响应的和。”

[2]Insofar as a load which has no magnetic or controlled source coupling to a one-port is concerned, a network containing linear elements and both independent and controlled sources may be replaced by an ideal voltage source of strength, \hat{V}_T , and an equivalent impedance Z_T , in series with the source. 句中“insofar as”短语意为“只要”，“one-port is concerned”在这里指的是等效电路的单端口。主句为“A network may be replaced by an ideal voltage source of strength, \hat{V}_T , and an equivalent impedance Z_T , in series with the source”;“containing linear elements and both independent and controlled sources”为现在分词短语作为定语,修饰主语“network”。此句可翻译为“当负载没有磁耦合或者受控电源耦合到等效电路的端口,网络仅仅包括线性元件和独立电压与受控电压时,该网络可以用一个理想电压源 \hat{V}_T 和一个等效阻抗 Z_T 串联来替代。”

1.3.4 Reference Translation

网络定理

线性性

如图 1.3.1 所示,我们用一个方框来表示某电路系统(也许仅包括一个网络元件),该系统输入用符号 e 来表示(表示激励),输出用符号 r 表示(表示响应)。如果此系统能满足齐次性和可加性条件,则称之为线性的。

齐次性条件:若系统的任意输入为 e 时,其响应为 r ,那么当输入为 ce 时,输出则为 cr ,其中 c 为常数。

可加性条件:若系统输入为 e_1 时,其响应为 r_1 ;系统输入为 e_2 时,响应为 r_2 ,那么系统输入为 $e_1 + e_2$ 时,输出则为 $r_1 + r_2$ 。

如果系统既不满足齐次性也不满足可加性,则我们称之为非线性系统。

叠加原理

线性网络的必要条件是齐次性和可加性,而可加性本身就是叠加原理的基础。

假设激励和响应是线性相关的,那么多个激励同时作用在系统上产生的总的响应等于每个激励分别单独作用时产生的响应的和。

例 1.3.1 在图 1.3.2(a)所示电路中,网络由左边电流源和上边的电压源驱动。电流相量 \hat{I} 待测。根据叠加原理,电流 \hat{I} 应为电流 \hat{I}_V 和电流 \hat{I}_C 之和,它们分别为电压源作用时产生的电流 \hat{I}_V ,如图 1.3.2(b)所示;和电流源单独作用时产生的电流 \hat{I}_C ,如图 1.3.2(c)所示。

$$\hat{I} = \hat{I}_V + \hat{I}_C$$

图 1.3.2(b)和(c)是采用去掉独立电压源和独立电流源的方法得到的。将独立电压源短路,独立电流源开路便可将它们从网络中去除。

图 1.3.2(b)和(c)所示为相量域中的简单梯形网络,求解方法为首先确定电压源和电流源的等效阻抗。在图 1.3.2(b)中,右边电压源的两个阻抗是串并联的,其阻抗为

$$Z_P = \frac{(40 - j40)(40j)}{40 + 40j - j40} = 40 + j40 \Omega$$

对电压源的总阻抗为

$$Z = Z_P + 40 - j40 = 40 + j40 + 40 - j40 = 80 \Omega$$

那么流出电压源的电流 \hat{I}_1 为

$$\hat{I}_1 = \frac{240 + j0}{80} = 3 + j0 \text{ A}$$

经过分流后

$$\hat{I}_V = \left[\frac{j40}{40 - j40 + j40} \right] (3 + j0) = j(3 + j0) = 0 + 3j \text{ A}$$

在图 1.3.2(b)中,电流源分别分流给 40Ω 电阻和一个由电容与 Z_P 组成的阻抗。我们称之为 Z_a ,则有

$$Z_a = -j40 + Z_p = -j40 + 40 + j40 = 40 \Omega$$

经两次分流后 \hat{I}_c 为

$$\hat{I}_c = \left[\frac{40}{40 + 40} \right] \left[\frac{j40}{40 - j40 + j40} \right] (0 + 6j) = \frac{j}{2} (0 + 6j) = 3 + 0j \text{ A}$$

图 1.3.2(a) 的电流 \hat{I} 为

$$\hat{I} = \hat{I}_v + \hat{I}_c = 0 + j3 + (3 + j0) = 3 + 3j \text{ A}$$

戴维南定理与诺顿定理

如果我们仅仅关心的是网络中的一部分的电压和电流,例如图 1.3.3(a) 的网络 B,而对另外一部分复杂电网络 A 没有兴趣,我们可以用一个简单的等值电路替换 A。这个简单的等效电路可由一个等效电压源和一个等效阻抗串联组成,如图 1.3.3(b) 所示。这种等效电路我们称之为戴维南等效电路。这种等效电路也可由一个等效电流源和一个阻抗并联来组成,如图 1.3.3(c) 所示,我们称之为诺顿等效电路。这里我们注意到只要 $Z_T = Z_N$ (下标 T 表示戴维南, N 表示诺顿),两种等效电路可以通过简单的电源变换来实现相互转换。

应用条件

在图 1.3.3(a) 中网络 A 的戴维南等效网络和诺顿等效网络仅仅在端口处是有效的,而不能延伸到电路内部。另外,对于网络 A 和 B 还有某些限制。网络 A 可能仅仅包括线性元件,但也可能包含独立电源或者受控源。而网络 B 并不局限于线性元件,它可以包括非线性元件或时变元件,或者独立的和非独立的电源。A 和 B 之间不能有受控的电源耦合或磁耦合。

戴维南定理

可以依据图 1.3.3(b) 来表述戴维南定理。

当负载没有磁耦合或者受控电源耦合到等效电路的端口,网络仅仅包括线性元件和独立电压与受控电压时,该网络可以用一个理想电压源 \hat{V}_T 和一个等效阻抗 Z_T 串联来替代。 \hat{V}_T 为网络端口两端的开路电压 \hat{V}_{oc} , Z_T 为网络端口的策动点的阻抗,它是把电路中所有独立电源置零后得到的。

诺顿定理

诺顿定理涉及的是电流源等效电路。可以依据图 1.3.3(c) 来描述此定理。

当负载没有磁耦合或者受控电源耦合到等效电路端口,网络仅仅包括线性元件和独立电压和受控电压时,该网络可以用一个理想电流源 \hat{I}_T 和一个等效阻抗 Z_N 并联来替代。 \hat{I}_T 是端口短路电流 \hat{I}_{sc} , Z_N 为网络端口的策动点阻抗,它是把电路中所有独立电源置零后得到的。

等效阻抗

有三种方法求解等效阻抗 Z_T ,使用哪种方法取决于电路分析人员的具体情况。不过,当有受控电源时,不能使用方法一。

方法一为直接计算 $Z_{eq} = Z_T = Z_N$,它是把电路中的所有独立电源全部置零后,从端口

看进去的等效阻抗,即将所有的独立电压源短路,独立电流源开路。

方法二可以用在含有受控源的电路中,需要计算戴维南等效电压(即网络端口的开路电压)和诺顿等效电流(即网络端口的短路电流)。其等效阻抗为这两个量的比

$$Z_T = Z_N = Z_{eq} = \frac{\hat{V}_T}{\hat{I}_N} = \frac{\hat{V}_{oc}}{\hat{I}_{sc}}$$

方法三也可以用在含有受控源的电路中。该方法需要在端口加一个测试电压,然后测量或者计算因此产生的电流。或者在端口串入测试电流,然后测量因此产生的端口电压。无论哪种方法,等值阻抗都可以由 \hat{V}_o 与 \hat{I}_o 的比值得到

$$Z_T = \frac{\hat{V}_o}{\hat{I}_o}$$

1.4 First-Order Circuits

1.4.1 Text

Introduction

Whenever a circuit is switched from one condition to another, either by a change in the applied source or a change in the circuit elements, there is a transitional period during which the branch currents and element voltages change from their former values to new ones. This period is called the transient. After the transient has passed, the circuit is said to be in the steady state. Now, the linear differential equation that describes the circuit will have two parts to its solution, the complementary function (or the homogeneous solution) and the particular solution. The complementary function corresponds to the transient, and the particular solution to the steady state.

In this chapter we will find the response of first-order circuits, given various initial conditions and sources. We will then develop an intuitive approach which can lead us to the same response without going through the formal solution of differential equations^[1]. We will also present and solve important issues relating to natural, force, step, and impulse responses, along with the dc steady state and the switching behavior of inductors and capacitors.

Capacitor Discharge in a Resistor

Assume a capacitor has a voltage difference V_o between its plates. When a conducting path R is provided, the stored charge travels through the capacitor from one plate to the other, establishing a current i . Thus, the capacitor voltage v is gradually reduced to zero, at which time the current also becomes zero. In the RC circuit of Figure 1.4.1(a), $Ri = v$ and $i = -Cd v/dt$. Eliminating i in both equations gives

$$\frac{dv}{dt} + \frac{1}{RC}v = 0 \quad (1.4.1)$$

The only function whose linear combination with its derivative can be zero is an exponential function of the form Ae^{st} . Replacing v by Ae^{st} and dv/dt by sAe^{st} in (1.4.1), we get

$$sAe^{st} + \frac{1}{RC}Ae^{st} = A(s + \frac{1}{RC})e^{st} = 0$$

from which

$$s + \frac{1}{RC} = 0 \quad \text{or} \quad s = -\frac{1}{RC} \tag{1.4.2}$$

Given $v(0) = A = V_0$ $v(t)$ and $i(t)$ are found to be

$$v(t) = V_0 e^{-t/RC}, t > 0 \tag{1.4.3}$$

$$i(t) = -C \frac{dv}{dt} = \frac{V_0}{R} e^{-t/RC}, t > 0 \tag{1.4.4}$$

The voltage and current of the capacitor are exponentials with initial values of V_0 and V_0/R , respectively. As time increases, voltage and current decrease to zero with a time constant of $\tau = RC$. See Figs. 1.4.1(b) and (c).

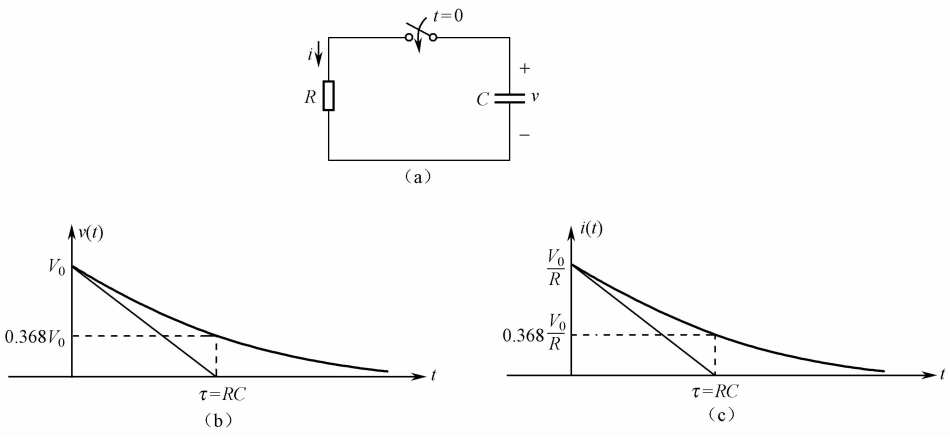


Figure 1.4.1 (a) A source-free RC circuit and its voltage, current responses.

Establishing a DC Voltage across a Capacitor

Connect an initially uncharged capacitor to a battery with voltage V_0 through a resistor at $t = 0$. The circuit is shown in Figure 1.4.2(a).

For $t > 0$, KVL around the loop gives $Ri + v = V_0$ which, after substituting $i = C(dv/dt)$, becomes

$$\frac{dv}{dt} + \frac{1}{RC}v = \frac{1}{RC}V_0, t > 0 \tag{1.4.5a}$$

with the initial condition

$$v(0^+) = v(0^-) = 0 \tag{1.4.5b}$$

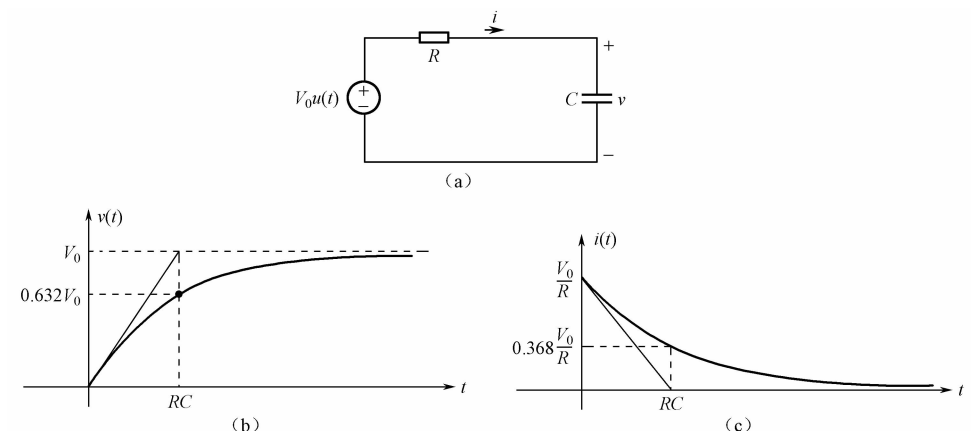


Figure 1.4.2 A RC circuit and its voltage, current responses

The solution should satisfy equations (1.4.5a) and (1.4.5b). The particular solution (or forced response) $v_p(t) = V_0$ satisfies equation (1.4.5a) but not equation (1.4.5b). The homogeneous solution (or natural response) $v_h(t) = Ae^{-t/RC}$ can be added and its magnitude A can be adjusted so that the total solution (1.4.6a) satisfies both (1.4.5a) and (1.4.5b).

$$v(t) = v_p(t) + v_h(t) = V_0 + Ae^{-t/RC} \quad (1.4.6a)$$

From the initial condition, $v(0^+) = V_0 + A = 0$ or $A = -V_0$. Thus the total solution is

$$v(t) = V_0(1 - e^{-t/RC})u(t) \quad (1.4.6b)$$

$$i(t) = \frac{V_0}{R}e^{-t/RC}u(t) \quad (1.4.6c)$$

The Source-Free RL Circuit

In the RL circuit of Figure 1.4.3, assume that at $t=0$ the current is I_0 . For $t > 0$, i should satisfy $Ri + L(di/dt) = 0$, the solution of which is $i = Ae^{st}$. By substitution we find A and s :

$$A(R + L_s)e^s = 0, \quad R + L_s = 0, \quad s = -R/L$$

The initial condition $i(0) = A = I_0$. Then

$$i(t) = I_0 e^{-Rt/L}, t > 0 \quad (1.4.7)$$

The time constant of the circuit is L/R .

Establishing a DC Current in an Inductor

If a dc source is suddenly applied to a series RL circuit initially at rest, as in Figure 1.4.4(a), the current grows exponentially from zero to a constant value with a time constant of L/R . The preceding result is the solution of the first-order differential equation (1.4.8) which is obtained by applying KVL around the loop. The solution follows.

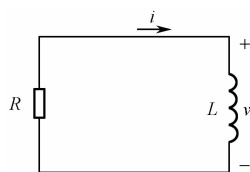


Figure 1.4.3 A source-free RL circuit.

$$Ri + L \frac{di}{dt} = V_0, \text{ for } t > 0, i(0^+) = 0 \quad (1.4.8)$$

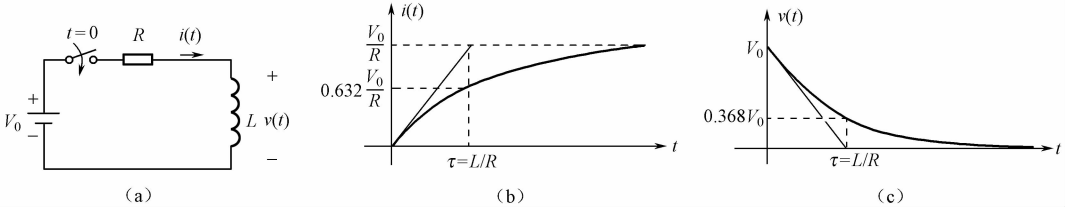


Figure 1.4.4 A RL circuit and its voltage, current responses.

Since $i = i_h(t) + i_p(t)$, where $i_h(t) = Ae^{-Rt/L}$ and $i_p(t) = v_0/R$, we have

$$i = Ae^{-Rt/L} + V_0/R$$

The coefficient A is found from $i(0^+) = A + V_0/R = 0$ or $A = -V_0/R$. The current in the inductor and the voltage across it are given by (1.4.9) and (1.4.10) and plotted in Figure 1.4.4(b) and (c), respectively.

$$i(t) = V_0/R(1 - e^{-Rt/L}), t > 0 \quad (1.4.9)$$

$$v(t) = L \frac{di}{dt} = V_0 e^{-Rt/L}, t > 0 \quad (1.4.10)$$

The Exponential Function Revisited

The exponential decay function may be written in the form $e^{-t/\tau}$, where τ is the time constant (in s). For RC the circuit shown in Figure 1.4.1(a), $\tau = RC$; while for RL the circuit shown in Figure 1.4.3, $\tau = L/R$. The general decay function

$$f = Ae^{-t/\tau}, t > 0$$

is plotted in Figure 1.4.5, with time measured in multiples of τ . It is seen that

$$f(\tau) = Ae^{-1} = 0.368A$$

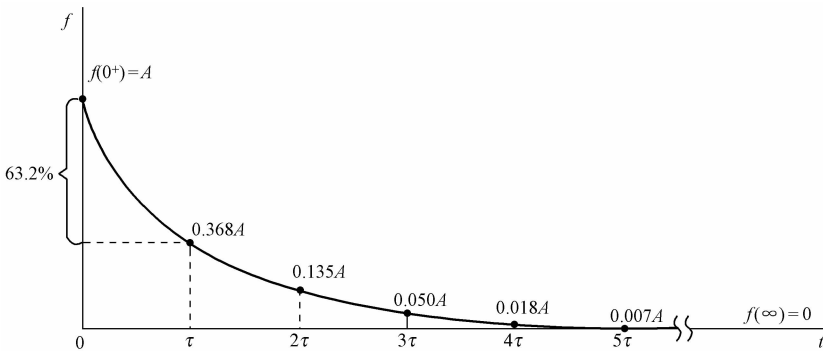


Figure 1.4.5 The plot of general decay function $f = e^{-t/\tau}$.

that is, at $t = \tau$ the function is 36.8 percent of the initial value. It may also be said that the function has undergone 63.2 percent of the change from $f(0^+)$ to $f(\infty)$. At $t = 5\tau$

the function has the value 0.0067, which is less than 1 percent of the initial value. From a practical standpoint, the transient is often regarded as over after $t = 5\tau$.

The tangent to the exponential curve at $t = 0^+$ can be used to estimate the time constant. In fact, since

$$\text{slope} = f'(0^+) = -\frac{A}{\tau}$$

the tangent line must cut the horizontal axis at $t = \tau$ (see Figure 1.4.6). More generally, the tangent at $t = t_0$ has horizontal intercept $t_0 + \tau$. Thus, if the two values $f(t_0)$ and $f'(t_0)$ are known, the entire curve can be constructed.

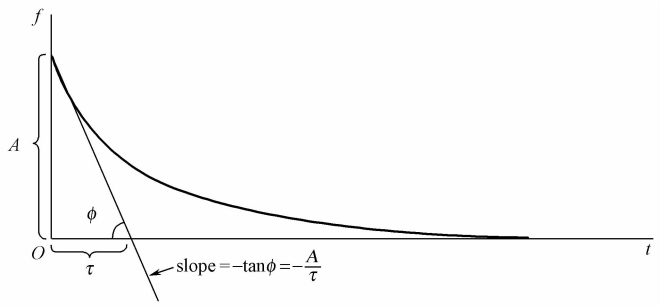


Figure 1.4.6 The tangent to the exponential curve at $t=0^+$.

At times a transient is only partially displayed (on chart paper or on the face of an oscilloscope), and the simultaneous values of function and slope needed in the preceding method are not available^[2]. In that case, any pair of data points, perhaps read from instruments, may be used to find the equation of the transient. Thus, referring to Figure 1.4.7,

$$f_1 = Ae^{-t_1/\tau}, f_2 = Ae^{-t_2/\tau}$$

which may be solved simultaneously to give

$$\tau = \frac{t_2 - t_1}{\ln f_1 - \ln f_2}$$

and then A in terms of τ and either f_1 or f_2 .

DC Steady State in Inductors and Capacitors

The natural exponential component of the response of RL and RC circuits to step inputs diminishes as time passes. At $t = \infty$, the circuit reaches steady state and the response is made of the forced dc component only.

Theoretically, it should take an infinite amount of time for RL or RC circuits to reach dc steady state. However, at $t = 5\tau$ the transient component is reduced to 0.67 percent of its initial value. After passage of 10 time constants the transient component equals to 0.0045 percent of its initial value, which is less than 5 in 100,000, at which time for all

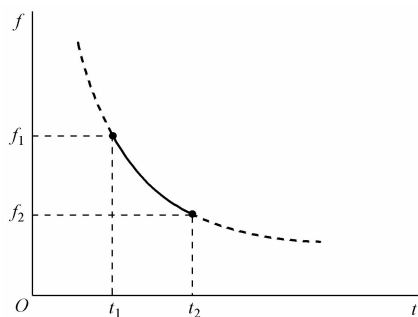


Figure 1.4.7 Two pairs of data points from instruments.

practical purposes we may assume the steady state has been reached^[3].

At the dc steady state of *RLC* circuits, assuming no sustained oscillations exist in the circuit, all currents and voltages in the circuit are constants. When the voltage across a capacitor is constant, the current through it is zero. All capacitors, therefore, appear as open circuits in

the dc steady state. Similarly, when the current through an inductor is constant, the voltage across it is zero. All inductors therefore appear as short circuits in the dc steady state. The circuit will be reduced to a dc-resistive case from which voltages across capacitors and currents through inductors can be easily found, as all the currents and voltages are constants and the analysis involves no differential equations^[4].

The dc steady-state behavior presented in the preceding paragraph is valid for circuits containing any number of inductors, capacitors, and dc sources.

1.4.2 Specialized English Words

first-order circuits 一阶电路

transient 暂态

complementary function 余函数

homogeneous solution 齐次解

particular solution 特解

initial conditions 初始条件

slope 斜率

oscilloscope 示波器

impulse response 脉冲电流

exponential function 指数函数

natural response 自然响应

tangent 正切

curve 曲线

decay 衰减

1.4.3 Notes

[1] We will then develop an intuitive approach which can lead us to the same response without going through the formal solution of differential equations. 句中的“develop”是在科技文献中使用频率较高的词,一般意思为“发展、开发或研究”,在这里根据上下文的意思,将原意引申译为“推导,导出”比较合适。此句可翻译为“无需通过正式的微分方程求解,我们将导出一种直观的方法来得到同样的响应。”

[2] At times a transient is only partially displayed (on chart paper or on the face of an oscilloscope), and the simultaneous values of function and slope needed in the preceding method are not available. 这里的“simultaneous”是指“function”和“slope”两者在同一时刻的值。此句可翻译为“有时暂态过程仅能显示一部分(在记录纸上或示波器

的屏幕上),而无法同时得到前面的方法所需的函数和斜率值。”

[3]After passage of 10 time constants the transient component equals to 0.0045 percent of its initial value, which is less than 5 in 100,000, at which time for all practical purposes we may assume the steady state has been reached. 句中“which is less than 5 in 100,000”是修饰“initial value”的定语从句,“at which time”中的“which”指的是时间点“After passage of 10 time constants”,“at which time”引导后面的时间状语从句。整句的主谓宾为“the transient component equals to 0.0045 percent of its initial value”。此句可翻译为“在经过 10 倍的时间常数后,暂态分量等于初始值的 0.0045%,小于十万分之五,此时对于实际应用来说,我们可以假定已经进入稳态。”

[4]The circuit will be reduced to a dc-resistive case from which voltages across capacitors and currents through inductors can be easily found, as all the currents and voltages are constants and the analysis involves no differential equations. 句中“all the current…”是“as”引导的原因状语从句,“from which voltages across… easily found”引导的是修饰主语“a dc-resistive case”的定语从句。此句可翻译为“当电路中所有的电压和电流均为常数且在分析中不涉及微分方程时,电路将简化为直流阻性电路,由此很容易求得电容端电压和流过电感的电流。”

1.4.4 Reference Translation

一 阶 电 路

介绍

无论何时,当电路中的电源或者电路元件发生改变时,电路都会由一种状态转换到另一种状态,在这个过程中,电路中的支路电流和元件上的电压的值也随之发生改变。这个过程称之为暂态过程。当过渡过程完成后,电路进入稳态。用来描述电路的线性微分方程有两部分解,一个是余函数(齐次解),另一个是特解。齐次解对应于暂态,特解对应于稳态。

本章我们将在各种不同的初始条件和电源的情况下求解一阶电路响应。无需通过正式的微分方程求解,我们将导出一种直观的方法来得到同样的响应。我们还将列出和解决有关自然响应、强制响应、阶跃响应和冲激响应的一些重要问题,包括直流稳态和电感电容的换路特性。

电容通过电阻放电

假定电容两极板间的电压差为 V_0 。当存在由电阻 R 构成的导电通路时,存储在电容上的电荷从一个极板向到另外一个极板运动,形成电流 i 。这样电容上的电压逐渐降到零,与此同时电流也逐渐降到零。在图 1.4.1(a)的 RC 电路中, $Ri = v$, $i = -Cd v/dt$ 。在方程中消去电流 i 得到

$$\frac{dv}{dt} + \frac{1}{RC}v = 0 \tag{1.4.1}$$

只有与其导数的线性组合为零的函数才是一个形式为 Ae^{st} 的指数函数。在式(1.4.1)中用 Ae^{st} 替换 v , 用 sAe^{st} 替代 dv/dt , 可以得到

$$sAe^{st} + \frac{1}{RC}Ae^{st} = A(s + \frac{1}{RC})e^{st} = 0$$

由此得

$$s + \frac{1}{RC} = 0 \quad \text{或} \quad s = -\frac{1}{RC} \quad (1.4.2)$$

给定 $v(0) = A = V_0$, 可求得 $v(t)$ 和 $i(t)$ 为

$$v(t) = V_0 e^{-t/RC}, t > 0 \quad (1.4.3)$$

$$i(t) = -C \frac{dv}{dt} = \frac{V_0}{R} e^{-t/RC}, t > 0 \quad (1.4.4)$$

电容上的电压和电流分别是初始值为 V_0 和 V_0/R 的指数。随着时间增长, 电压和电流以 $\tau = RC$ 为时间常数减小到零。见图 1.4.1(b)和(c)。

在电容两端建立直流电压

在 $t=0$ 时刻, 将电压 V_0 接入到初始值为零电容和电阻串联的回路中。电路如图 1.4.2(a)所示。

当 $t>0$, 沿回路应用 KVL 定理得到 $Ri + v = V_0$, 将 $i = C(dv/dt)$ 代入, 得到

$$\frac{dv}{dt} + \frac{1}{RC}v = \frac{1}{RC}V_0, t > 0 \quad (1.4.5a)$$

初始条件为

$$v(0^+) = v(0^-) = 0 \quad (1.4.5b)$$

答案应同时满足式(1.4.5a)和式(1.4.5b)。特解(或强迫响应) $v_p(t) = V_0$ 满足式(1.4.5a)但不满足式(1.4.5b)。加齐次解(或自然响应) $v_h(t) = Ae^{-t/RC}$ 并调整其幅值 A 使得全解式(1.4.6a)满足式(1.4.5a)和式(1.4.5b)。

$$v(t) = v_p(t) + v_h(t) = V_0 + Ae^{-t/RC} \quad (1.4.6a)$$

根据初始条件, $v(0^+) = V_0 + A = 0$ 或 $A = -V_0$ 。全解为

$$v(t) = V_0(1 - e^{-t/RC})u(t) \quad (1.4.6b)$$

$$i(t) = \frac{V_0}{R} e^{-t/RC} u(t) \quad (1.4.6c)$$

不含电源的 RL 电路

如图 1.4.3 所示 RL 电路中, 假设在 $t = 0$ 时电流为 I_0 。 $t > 0$ 时 i 应该满足 $Ri + L(di/dt) = 0$, 它的解是 $i = Ae^{st}$ 。用代入法可求得 A 和 s

$$A(R + Ls)e^{st} = 0, \quad R + Ls = 0, \quad s = -R/L$$

初始条件 $i(0) = A = I_0$, 则有

$$i(t) = I_0 e^{-Rt/L}, t > 0 \quad (1.4.7)$$

电路的时间常数是 L/R 。

在电感中建立一个直流电流

如图 1.4.4(a)所示, 如果一个直流电源突然加到一个初始值为零的串联 RL 电路, 电

流将从零以指数形式增长到一个定值,时间常数为 L/R 。以上结果为一阶微分方程(1.4.8)的解,这个可以对回路应用 KVL 求得。解答如下。

$$Ri + L \frac{di}{dt} = V_0, \text{对于 } t > 0, i(0^+) = 0 \quad (1.4.8)$$

由于 $i = i_h(t) + i_p(t)$, 这里

$$i_h(t) = Ae^{-Rt/L}, i_p(t) = v_0/R$$

可得

$$i = Ae^{-Rt/L} + V_0/R$$

系数 A 从 $i(0^+) = A + V_0/R = 0$ 或 $A = -V_0/R$ 求得。电感中的电流和它两端的电压由式(1.4.9)和式(1.4.10)得出并分别画于图 1.4.7(b)和(c)中。

$$i(t) = V_0/R(1 - e^{-Rt/L}), t > 0 \quad (1.4.9)$$

$$v(t) = L \frac{di}{dt} = V_0 e^{-Rt/L}, t > 0 \quad (1.4.10)$$

再谈指数函数

指数衰减函数可以用 $e^{-t/\tau}$ 的形式来表示,这里 τ 是时间常数(单位为秒)。对于图 1.4.1(a)所示的 RC 电路, $\tau = RC$; 而对于图 1.4.3 所示的 RL 电路, $\tau = L/R$ 。一般衰减函数为

$$f = Ae^{-t/\tau}, t > 0$$

将其画于图 1.4.5 中,时间用 τ 的倍数表示。可以看到

$$f(\tau) = Ae^{-1} = 0.368A$$

也就是说,在 $t = \tau$ 时,函数值为初始值的 36.8%。也可以说函数值经历了从 $f(0^+)$ 到 $f(\infty)$ 的 63.2% 的变化。在 $t = 5\tau$ 处,函数值为 0.0067A,小于初始值的 1%。从实践的观点来看,可以认为暂态过程在 $t = 5\tau$ 之后就结束了。

我们可以用指数曲线在 $t = 0^+$ 时的切线来估计时间常数。实际上,由于

$$\text{斜率} = f'(0^+) = -\frac{A}{\tau}$$

正切线必定在 $t = \tau$ 时与横坐标轴相交(见图 1.4.6)。更一般的情况为,正切线在 $t = t_0$ 时与横轴交于 $t_0 + \tau$ 。这样,如果 $f(t_0)$ 和 $f'(t_0)$ 的值已知,就可以画出整个曲线。

有时暂态过程仅能显示一部分(在记录纸上或示波器的屏幕上),而无法同时得到前面的方法所需的函数和斜率值。在这种情况下,从仪器上读取的任意一对数据点的值可以用来求得暂态方程。参见图 1.4.7,

$$f_1 = Ae^{-t_1/\tau}, \quad f_2 = Ae^{-t_2/\tau}$$

联立求解可以得到

$$\tau = \frac{t_2 - t_1}{\ln f_1 - \ln f_2}$$

再根据 τ 和 f_1 或者 f_2 可以得 A 。

电感电容的直流稳态

RL 和 RC 电路对于阶跃输入的响应的自然指数分量随时间减小。在 $t = \infty$ 时, 电路进入稳态, 电路响应只剩下强制直流分量。

理论上, 对于 RL 和 RC 电路应该需要无限长时间才能进入直流稳态。然而, 在 $t = 5\tau$ 时, 暂态分量减少到初始值的 0.67% 。在经过 10 倍的时间常数后, 暂态分量等于初始值的 0.0045% , 小于十万分之五, 此时对于实际应用来说, 我们可以假定已经进入稳态。

假设电路中不存在持续的振荡, RLC 电路处在的直流稳态状态时, 所有的电压和电流都是常数。当电容两端的电压恒定不变时, 通过它的电流则为零。因此在直流稳态电路中, 所有的电容均为开路。类似地, 当流过电感的电流恒定不变, 则它两端的电压为零。因此在直流稳态电路中, 所有电感看起来如同短路一样。当电路中所有的电压和电流均为常数且在分析中不涉及微分方程时, 电路将简化为直流阻性电路, 由此很容易求得电容端电压和流过电感的电流。

上述直流稳态的特性适用于包含任意数量的电感、电容和直流电源的电路。

1.5 Sinusoidal Steady-State Circuit Analysis

1.5.1 Text

Introduction

This chapter will concentrate on the steady-state response of circuits driven by sinusoidal sources. The response will also be sinusoidal. For a linear circuit, the assumption of a sinusoidal source represents no real restriction, since a source that can be described by a periodic function can be replaced by an equivalent combination (Fourier series) of sinusoids.

Element Responses

The voltage-current relationships for the single elements R , L , and C were summarized in Table 1.5.1. In this chapter, the functions of v and i will be sines or cosines with the argument ωt . ω is the angular frequency and has the unit rad/s. Also, $\omega = 2\pi f$, where f is the frequency with unit cycle/s, or more commonly hertz (Hz).

Consider an inductance L with $i = I\cos(\omega t + 45^\circ)$ A [see Figure 1.5.1(a)]. The voltage is

$$v_L = L \frac{di}{dt} = \omega L I [-\sin(\omega t + 45^\circ)] = \omega L I \cos(\omega t + 135^\circ) \text{ (V)}$$

A comparison of v_L and i shows that the current lags the voltage by 90° or $\pi/2$ rad. The functions are sketched in Figure 1.5.1(b). Note that the current function i is to the right of v , and since the horizontal scale is ωt , events displaced to the right occur later in time. This illustrates that i lags v . The horizontal scale is in radians, but note that it

is also marked in degrees ($-135^\circ, 180^\circ$, etc.). This is a case of mixed units just as with $\omega t + 45^\circ$. It is not mathematically correct but is the accepted practice in circuit analysis. The vertical scale indicates two different quantities, that is, v and i , so there should be two scales rather than one.

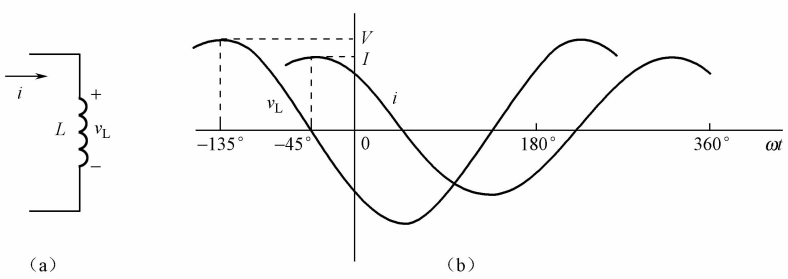
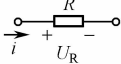
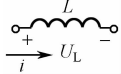
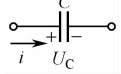


Figure 1.5.1 The voltage and current waveforms in an inductance L .

While examining this sketch, it is a good time to point out that a sinusoid is completely defined when its magnitude (V or I), frequency (ω or f), and phase (45° or 135°) are specified^[1].

In Table 1.5.1 the responses of the three basic circuit elements are shown for applied current $i = I \cos \omega t$ and voltage $v = V \cos \omega t$. If sketches are made of these responses, they will show that for a resistance R , v and i are in phase. For an inductance L , i lags v by 90° or $\pi/2$ rad. And for a capacitance C , i leads v by 90° or $\pi/2$ rad.

Table 1.5.1 The voltage-current relationships of the three basic circuit elements.

	$i = I \cos \omega t$	$v = V \cos \omega t$
	$v_R = RI \cos \omega t$	$i_R = R \cos \omega t$
	$v_L = \omega L I \cos(\omega t + 90^\circ)$	$i_L = \frac{V}{\omega L} \cos(\omega t - 90^\circ)$
	$v_C = \frac{I}{\omega C} \cos(\omega t - 90^\circ)$	$i_C = \omega C V \cos(\omega t + 90^\circ)$

Phasors

A brief look at the voltage and current sinusoids in the preceding examples shows that the amplitudes and phase differences are the two principal concerns^[2]. A directed line segment, or phasor, such as that shown rotating in a counterclockwise direction at a constant angular velocity ω (rad/s) in Figure 1.5.2, has a projection on the horizontal which is a cosine function^[3]. The length of the phasor or its magnitude is the amplitude or maximum value of the cosine function. The angle between two positions of the phasor

is the phase difference between the corresponding points on the cosine function.

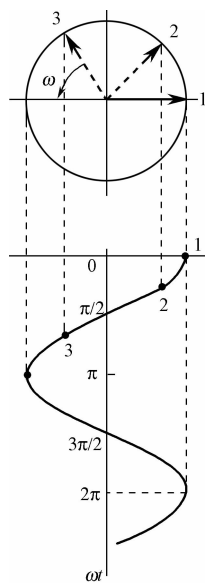


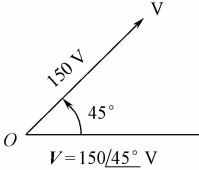
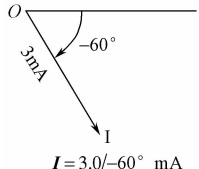
Figure 1. 5. 2 A phasor rotating counterclockwise an its projectin on the real axis.

Throughout this book phasors will be defined from the cosine function. If a voltage or current is expressed as a sine, it will be changed to a cosine by subtracting 90° from the phase.

Consider the examples shown in Table 1. 5. 2. Observe that the phasors, which are directed line segments and vectorial in nature, are indicated by boldface capitals, for example, **V**, **I**. The phase angle of the cosine function is the angle on the phasor. The phasor diagrams here and all that follow may be considered as a snapshot of the counterclock- wise- rotating directed line segment taken at $t = 0^{[4]}$. The frequency $f(\text{Hz})$ and $\omega(\text{rad/s})$ generally do not appear but they should be

kept in mind, since they are implicit in any sinusoidal steady-state problem.

Table 1. 5. 2 Functions and its phasor representation.

Function	Phasor Representation
$v=150\cos(500t+45^{\circ})\text{ (V)}$	
$i=3.0\sin(2000t+30^{\circ})\text{ (mA)}$ $=3.0\cos(2000t-60^{\circ})\text{ (mA)}$	

Impedance and Admittance

A sinusoidal voltage or current applied to a passive RLC circuit produces a sinusoidal response. With time functions, such as $v(t)$ and $i(t)$, the circuit is said to be in the time domain, Figure 1. 5. 3(a); and when the circuit is analyzed using phasors, it is said to be in the frequency domain, Figure 1. 5. 3(b). The voltage and current may be

written, respectively,

$$v(t) = V\cos(\omega t + \varphi) = \text{Re}[Ve^{j\omega t}], V = V\angle\varphi$$

$$i(t) = I\cos(\omega t + \varphi) = \text{Re}[Ie^{j\omega t}], I = I\angle\varphi$$

The ratio of phasor voltage V to phasor current I is defined as impedance Z , that is, $Z = V/I$. The reciprocal of impedance is called admittance Y , so that $Y = 1/Z$ (S), where $1 \text{ S} = 1 \Omega^{-1} = 1 \text{ mho}$. Y and Z are complex numbers.

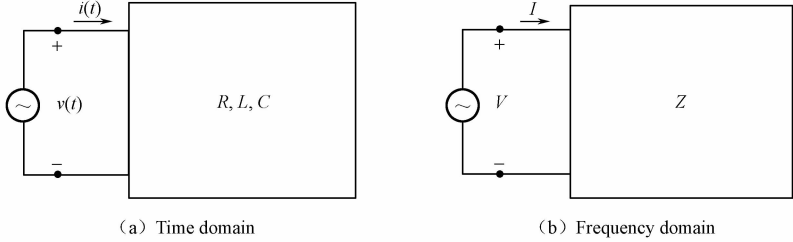


Figure 1.5.3 (a) Time domain and (b) frequency domain circuits.

When impedance is written in Cartesian form the real part is the resistance R and the imaginary part is the reactance X . The sign on the imaginary part may be positive or negative: When positive, X is called the inductive reactance, and when negative, X is called the capacitive reactance. When the admittance is written in Cartesian form, the real part is admittance G and the imaginary part is susceptance B . A positive sign on the susceptance indicates a capacitive susceptance, and a negative sign indicates an inductive susceptance. Thus,

$$Z = R + jX_L \text{ and } Z = R - jX_C$$

$$Y = G - jB_L \text{ and } Y = G + jB_C$$

The relationships between these terms follow from $Z = 1/Y$. Then,

$$R = \frac{G}{G^2 + B^2} \quad \text{and} \quad X = \frac{-B}{G^2 + B^2}$$

$$G = \frac{R}{R^2 + X^2} \quad \text{and} \quad B = \frac{-X}{R^2 + X^2}$$

These expressions are not of much use in a problem where calculations can be carried out with the numerical values as in the following example.

Combinations of Impedances The relation $V = IZ$ (in the frequency domain) is formally identical to Ohm's law, $v = iR$, for a resistive network (in the time domain). Therefore, impedances combine exactly like resistances:

impedances in series $Z_{eq} = Z_1 + Z_2 + \dots$

impedances in parallel $\frac{1}{Z_{eq}} = \frac{1}{Z_1} + \frac{1}{Z_2} + \dots$

In particular, for two parallel impedances, $Z_{eq} = Z_1 Z_2 / (Z_1 + Z_2)$.

Impedance Diagram In an impedance diagram, an impedance Z is represented by a point in the right half of the complex plane. Figure 1.5.4 shows two impedances, Z_1 , in the first quadrant, exhibits inductive reactance, while Z_2 , in the fourth quadrant, exhibits capacitive reactance. Their series equivalent, $Z_1 + Z_2$, is obtained by vector addition, as shown. Note that the “vectors” are shown without arrowheads, in order to distinguish these complex numbers from phasors.

Combinations of Admittances Replacing Z by $1/Y$ in the formulas above gives

admittances in series $\frac{1}{Y_{eq}} = \frac{1}{Y_1} + \frac{1}{Y_2} + \cdots$

admittances in parallel $Y_{eq} = Y_1 + Y_2 + \cdots$

Thus, series circuits are easiest treated in terms of impedance; parallel circuits, in terms of admittance.

Admittance Diagram Figure 1.5.5, an admittance diagram, is analogous to Figure 1.5.4 for impedance. Shown are an admittance Y_1 having capacitive susceptance and an admittance Y_2 having inductive susceptance, together with their vector sum, $Y_1 + Y_2$, which is the admittance of a parallel combination of Y_1 and Y_2 .

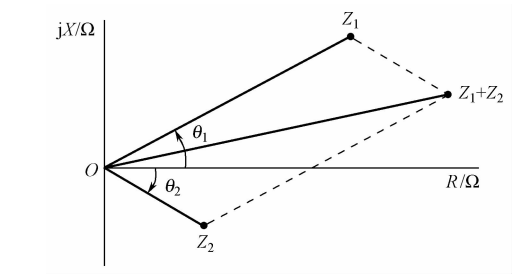


Figure 1.5.4 An impedance diagram.

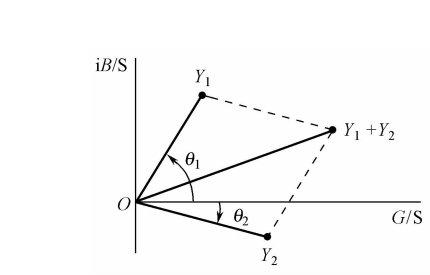


Figure 1.5.5 A admittance diagram.

1.5.2 Specialized English Words

steady-state response 稳态响应
 sinusoidal sources 正弦电源
 Fourier series 傅里叶级数
 sines 正弦
 cosines 余弦
 argument 自变量,参数
 hertz 赫兹
 lag 滞后
 horizontal scale 水平刻度
 radians 弧度

degree 度
 units 单位
 vertical scale 垂直刻度
 magnitude 大小,幅值
 in phase 同相
 phasor 相量
 counterclockwise 逆时针
 angular velocity 角速度
 projection 投影
 function 函数

amplitude 幅值	inductive reactance 感抗
boldface 黑体	capacitive reactance 容抗
admittance 导纳	susceptance 电纳
time domain 时域	capacitive susceptance 容纳
frequency domain 频域	inductive susceptance 感纳
Cartesian form 笛卡儿坐标系, 直角坐标系	quadrant 象限
real part 实部	complex plane 复平面
imaginary part 虚部	vector 矢量, 向量
reactance 电抗	

1. 5. 3 Notes

[1] While examining this sketch, it is a good time to point out that a sinusoid is completely defined when its magnitude (V or I), frequency (ω or f), and phase (45° or 135°) are specified. 句中“Examining”如果翻译为检查,显得过于生硬,在这里应该翻译为“研究、分析”。“It is a good time”可以省略不翻。此句可以翻译为“分析此图会发现,一个具有完整定义的正弦波应当指明其幅值(V 或 I)、频率(ω 或 f)和相位(45° 或 135°)。”

[2] A brief look at the voltage and current sinusoids in the preceding examples shows that the amplitudes and phase differences are the two principal concerns. 句中的“a brief look…”这里表示简单地看看,“that the amplitudes and phase…”是“that”引导的宾语从句,其中“concerns”表示关切的事情。此句可以翻译为“简单分析前面例子中所示的电压和电流正弦曲线,就会看出幅值差和相位差是值得关注的两个主要问题。”

[3] A directed line segment, or phasor, such as that shown rotating in a counterclockwise direction at a constant angular velocity ω (rad/s) in Figure 1. 5. 2, has a projection on the horizontal which is a cosine function. 句中的“such as that shown”省略了“which is, shown…”,在这里作为过去分词结构作为“that”的后置定语。此句可以翻译为“如图 1. 5. 2 中以恒定的角速度 ω (rad/s) 按逆时针方向旋转的相量,一条有向的线段,或者相量在横坐标轴上的投影为余弦函数。”

[4] The phasor diagrams here and all that follow may be considered as a snapshot of the counter clockwise-rotating directed line segment taken at $t=0$. 句中的“that follow”是“all”的定语从句,指的是相量从角度为零的地方开始逆时针旋转后,随后的那些相量。此句可以翻译为“在这里的相量图和其后的相量图可以被认为是被认定为有向线段在 $t=0$ 时刻按逆时针方向旋转的快照。”

1. 5. 4 Reference Translation

正弦稳态电路分析

引言

本章将主要讲述由正弦电源驱动的电路的稳态响应。其响应也是正弦形式。对于一

个线性电路来说,它的信号源不一定真是正弦信号源,因为一个周期性函数的电源信号,可以被一个等效的正弦函数组合(傅里叶级数)所代替。

元件响应

表 1.5.1 中总结了作为单独元件的 R 、 L 和 C 的电压-电流关系。在本文中,函数 v 和 i 将会是以 ωt 为变量的正弦或余弦函数, ω 是角频率,单位为 rad/s ,另外 $\omega = 2\pi f$, 其中 f 是频率,单位是周期/s,或更一般地为赫兹(Hz)。

考虑电感 L ,其电流 $i = I\cos(\omega t + 45^\circ)$ A [见图 1.5.1(a)]。电压为

$$v_L = L \frac{di}{dt} = \omega L I [-\sin(\omega t + 45^\circ)] = \omega L I \cos(\omega t + 135^\circ) (\text{V})$$

比较曲线 v_L 和曲线 i ,表明在相位上,电流滞后电压 90° 或者说滞后 $\pi/2\text{rad}$ 。这个函数画于图 1.5.1(b)中。注意电流函数 i 在电压函数 v 的右侧,由于横坐标为 ωt ,所以右侧的事件在时间上要晚一些发生。这说明电流滞后于电压。横坐标单位是弧度,但也可以用度数来表示(-135° 、 180° 等)。所以这是一个混合单位的情况,如 $\omega t + 45^\circ$ 。这在数学上是不正确的,但在电路分析中是实际上可接受的。纵坐标代表两个不同的量,即 v 和 i ,所以应有两个刻度而不是一个。

分析此图会发现,一个具有完整定义的正弦波应当指明其幅值(V 或 I)、频率(ω 或 f)和相位(45° 或 135°)。

表 1.5.1 给出了在施加电流 $i = I\cos\omega t$ 和电压 $v = V\cos\omega t$ 时,三种基本电路元件的响应情况。如果画出这些响应的图形,则可以看到对于电阻 R , v 和 i 是同相位的。对于电感 L , i 滞后于 v 90° 或者 $\pi/2 \text{ rad}$ 。对于电容 C , i 超前 v 90° 或者 $\pi/2 \text{ rad}$ 。

相量

简单分析前面例子中所示的电压和电流正弦曲线,就会看出幅值差和相位差是值得关注的两个主要问题。如图 1.5.2 中以恒定的角速度 $\omega(\text{rad/s})$ 按逆时针方向旋转的相量,一条有向的线段,或者相量在横坐标轴上的投影为余弦函数。相量的长度或者幅值是余弦函数的幅值或最大值。相量的两个位置之间的角度是余弦函数上相应的点之间的相位差。

本书中全部用余弦函数对相量进行定义。如果电压或电流是用正弦函数表示的,则采用将其相位减去 90° 的方法,将其转化为余弦函数。

下面看表 1.5.2 中的几个例子。我们注意到实质上是有向线段和矢量的这些相量是用大写黑体字母来标注的,如 \mathbf{V} 、 \mathbf{I} 。余弦函数的相位角即为相量的角度。在这里的相量图和其后的相量图可以被认为是 有向线段在 $t=0$ 时刻按逆时针方向旋转的快照。频率 $f(\text{Hz})$ 和 $\omega(\text{rad/s})$ 一般不在图中标出,由于它们一般隐含在所有正弦稳态电路的问题中,因此应该记住它们。

阻抗和导纳

正弦电压或电流加到无源 RLC 电路中会产生正弦响应。对于采用如同 $v(t)$ 和 $i(t)$ 的时间函数,电路称时域电路,如图 1.5.3(a)所示;当采用相量来分析电路时,我们称电

路为频域电路,如图 1.5.3(b)所示。相应电压和电流可以分别写为

$$\begin{aligned} v(t) &= V\cos(\omega t + \varphi) = \operatorname{Re}[Ve^{j\omega t}], V = V\angle\varphi \\ i(t) &= I\cos(\omega t + \varphi) = \operatorname{Re}[Ie^{j\omega t}], I = I\angle\varphi \end{aligned}$$

电压相量 V 和电流相量 I 的比值定义为阻抗 Z , 即 $Z=V/I$ 。阻抗的倒数称为导纳 $Y, Y=1/Z(S)$, 这里 $1S = 1\ \Omega^{-1} = 1$ 姆欧。其中 Y 和 Z 均为复数。

当阻抗写成直角坐标形式时,实部是电阻 R ,虚部是电抗 X 。虚部的符号可正可负:如果为正, X 称为感抗;如果为负, X 称为容抗。当导纳写成直角坐标形式时,其实部是导纳 G ,虚部是电纳 B 。电纳的符号也可正可负,如为正,则称为容纳,如为负,则为感纳。因此

$$\begin{aligned} Z &= R + jX_L & Z &= R - jX_C \\ Y &= G - jB_L & Y &= G + jB_C \end{aligned}$$

这些项之间的关系均遵循公式 $Z=1/Y$ 。则

$$\begin{aligned} R &= \frac{G}{G^2 + B^2} \text{ 和 } X = \frac{-B}{G^2 + B^2} \\ G &= \frac{R}{R^2 + X^2} \text{ 和 } B = \frac{-X}{R^2 + X^2} \end{aligned}$$

这些表达式在问题中用的并不多,如同下面的例子一样,可采用数值方法进行计算。

阻抗的合并 对于一个电阻网络, $V = IZ$ (频域)在形式上等同于欧姆定律, $v = iR$ (时域)。因此,阻抗可像电阻那样进行合并。

$$\begin{aligned} \text{串联阻抗} \quad Z_{eq} &= Z_1 + Z_2 + \cdots \\ \text{并联阻抗} \quad \frac{1}{Z_{eq}} &= \frac{1}{Z_1} + \frac{1}{Z_2} + \cdots \end{aligned}$$

特别说明,对于两个并联的阻抗,

$$Z_{eq} = Z_1 Z_2 / (Z_1 + Z_2)$$

阻抗图 在阻抗图中,阻抗 Z 是用复平面右半平面中一个点来表示的。图 1.5.4 所示两个阻抗: Z_1 在第一象限,表示它是一个感抗,而 Z_2 在第四象限,说明它是一个容抗。如图 1.5.4 所示,它们的串联等效值 $Z_1 + Z_2$,可用向量加法求得。注意,图中的向量没有箭头,这是为了将这些复数与相量区别开。

导纳合并 将上面公式中的 Z 用 $1/Y$ 替换可以得到

$$\begin{aligned} \text{串联导纳} \quad \frac{1}{Y_{eq}} &= \frac{1}{Y_1} + \frac{1}{Y_2} + \cdots \\ \text{并联导纳} \quad Y_{eq} &= Y_1 + Y_2 + \cdots \end{aligned}$$

这样,串联电路用阻抗最容易处理,而并联电路用导纳最容易处理。

导纳图 如图 1.5.5 所示的导纳图,与图 1.5.4 所示的阻抗图类似。图中导纳 Y_1 有容纳,导纳 Y_2 有感纳,它们的向量和 $Y_1 + Y_2$ 是 Y_1 和 Y_2 的并联组合的导纳。

Part 2 Electronics

2.1 Interpreting a Digital IC Datasheet

2.1.1 Text

Semiconductor manufacturers publish data sheets for each of their products. Regardless of the specific family or device, all logic IC data sheets share common types of information. Once the basic data sheet terminology and organization is understood, it is relatively easy to figure out other data sheets even when their exact terminology changes^[1]. Data sheet structure is illustrated using the 74LS00 from Fairchild Semiconductor as an example. A page from its data sheet is shown in Figure 2.1.1.

Digital IC data sheets should have at least two major sections: functional description and electrical specifications. The functional description usually contains the device pin assignment, or pin-out, as well as a detailed discussion of how the part logically operates. A simple IC such as the 74LS00 will have a very brief functional description, because there is not much to say about a NAND gate's operation. More complex ICs such as microprocessors can have functional descriptions that fill dozens or hundreds of pages and are broken into many chapters. Some data sheets add additional sections to present the mechanical dimensions of the package and its thermal properties. Digital IC electrical specifications are similar across most types of devices and often appear in the following four categories:

Absolute maximum ratings. As the term implies, these parameters specify the absolute extremes that the IC may be subjected to without sustaining permanent damage. Manufacturers almost universally state that the IC should never be operated under these extreme conditions. These ratings are useful, because they indicate how the device may be stored and express the quality of design and manufacture of the physical chip. Manufacturers specify a storage temperature range within which the semiconductor structures will not break down^[2]. In the case of Fairchild's 74LS00, this range is $-65\text{ }^{\circ}\text{C}$ to $150\text{ }^{\circ}\text{C}$. Maximum voltage levels are also specified, 7 V in the case of the 74LS00, indicating that the device may be subjected to a 7 V potential without destructing.

Recommended operating conditions. These parameters specify the normal range of voltages and temperatures that the IC should be operated within such that its functionality is guaranteed to meet specifications set forth by the manufacturer. Two of the most important specifications in this section are the supply voltage (commonly

Absolute Maximum Ratings(Note 1)

Supply Voltage	7V	Note 1: The "Absolute Maximum Ratings" are those values beyond which the safety of the device cannot be guaranteed. The device should not be operated at these limits. The parametric values defined in the Electrical Characteristics tables are not guaranteed at the absolute maximum ratings. The "Recommended Operating Conditions" table will define the conditions for actual device operation.
Input Voltage	7V	
Operating Free Air Temperature Range	0°C to +70°C	
Storage Temperature Range	-65°C to +150°C	

Recommended Operating Conditions

Symbol	Parameter	Min	Nom	Max	Units
V _{CC}	Supply Voltage	4.75	5	5.25	V
V _{IH}	HIGH Level Input Voltage	2			V
V _{IL}	LOW Level Input Voltage			0.8	V
I _{OH}	HIGH Level Output Current			-0.4	mA
I _{OL}	LOW Level Output Current			8	mA
T _A	Free Air Operating Temperature	0		70	°C

Electrical Characteristics

over recommended operating free air temperature range(unless otherwise noted)

Symbol	Parameter	Conditions	Min	Typ (Note 2)	Max	Units
V _I	Input Clamp Voltage	V _{CC} =Min, I _I =-18mA			-1.5	V
V _{OH}	HIGH Level Output Voltage	V _{CC} =Min, I _{OH} =Max, V _{IL} =Max	2.7	3.4		V
V _{OL}	LOW Level Output Voltage	V _{CC} =Min, I _{OL} =Max, V _{IH} =Min		0.35	0.5	V
		I _{OL} =4mA, V _{CC} =Min		0.25	0.4	
I _I	Input Current @ Max Input Voltage	V _{CC} =Max, V _I =7V			0.1	mA
I _{IH}	HIGH Level Input Current	V _{CC} =Max, V _I =2.7V			20	µA
I _{IL}	LOW Level Input Current	V _{CC} =Max, V _I =0.4V			-0.36	mA
I _{OS}	Short Circuit Output Current	V _{CC} =Max(Note 3)	-20		-100	mA
I _{CCH}	Supply Current with Outputs HIGH	V _{CC} =Max		0.8	1.6	mA
I _{CCL}	Supply Current with Outputs LOW	V _{CC} =Max		2.4	4.4	mA

Note 2: All typicats are at V_{CC}=5V, T_A=25°C

Note 3: Not more than one output should be shorted at a time, and the duration should not exceed one second.

Switching Characteristics

at V_{CC}=5V and T_A=25°C

Symbol	Parameter	$R_L = 2\text{ k}\Omega$				Units
		$C_L = 15\text{pF}$		$C_L = 50\text{pF}$		
		Min	Max	Min	Max	
t_{PLH}	Propagation Delay Time LOW-to-HIGH Level Output	3	10	4	15	ns
t_{PHL}	Propagation Delay Time HIGH-to-LOW Level Output	3	10	4	15	ns

Figure 2. 1. 1 74LS00 Datasheet.

labeled as either V_{CC} or V_{DD} , depending on whether a bipolar or MOS process) and the operating temperature. An IC may have multiple supply voltage specifications, because an IC can actually operate on several different voltages simultaneously. Each supply voltage may power a different portion of the chip. When the manufacturer specifies supply voltage, it does so with a certain tolerance, usually either ± 5 or ± 10 percent. Many 5 V logic ICs are guaranteed to operate only at a supply voltage from 4.75 to 5.25 V (± 5 percent). Operating temperature is very important, because it affects the timing of the device. As a semiconductor heats up, it slows down. As it cools, its speed increases. Outside of the recommended operating temperature, the device is not guaranteed to function, because the effects of temperature become so severe that functionality is compromised. There are four common temperature ranges for ICs: commercial (0°C to 70°C), industrial (-40°C to 85°C), automotive (-40°C to 125°C), and military (-55°C to 125°C). It is more difficult to manufacture an IC that operates over wider temperature ranges. As such, more demanding temperature grades are often more expensive than the commercial grade.

Other parameters establish the safe operating limits for input signals as well as the applied voltage thresholds that represent logic 0 and 1 states^[3]. Minimum and maximum input levels are expressed as either absolute voltages or voltages relative to the supply voltage pins of the device. Exceeding these voltages may damage the device. Logic threshold specifications are provided to ensure that the logic input voltages are such that the device will function as intended and not confuse a 1 for a 0, or vice versa. There is also a limit to how much current a digital output can drive. Current output specifications should be known so that a chip is not overloaded, which could result in either permanent damage to the chip or the chip's failure to meet its published specifications.

DC electrical characteristics. DC parameters specify the voltages and currents that the IC will present to other circuitry to which it is connected^[4]. Whereas recommended operating conditions specify the environment under which the chip will properly operate, DC electrical characteristics specify the environment that the chip itself will create. Output voltage specifications define the logic 0 and 1 thresholds that the chip is guaranteed to drive under all legal operating conditions. These specifications confirm that the chip is compatible with other chips in the same family and also allow an engineer to determine if the output levels are compatible with another chip that it may be driving.

Input current specifications characterize the load that the chip presents to whatever circuit is driving it. When either logic state is applied to the chip, a small current flows between the driver and the chip in question. Quantifying these currents enables an engineer to ensure compatibility between multiple ICs. When one IC drives several other ICs, the sum of the input currents should not exceed the output current specification of the driver.

AC electrical characteristics (or switching characteristics). AC parameters often represent the greatest complexity and level of detail in a digital IC's specifications. They are the guaranteed timing parameters of inputs and outputs. If the IC is purely combinatorial (e. g. , 74LS00), timing may just be matter of specifying propagation delays and rise and fall times. Logic ICs with synchronous elements (e. g. , flops) have associated parameters such as setup, hold, clock frequency, and output valid times.

Keep in mind that each manufacturer has a somewhat different style of presenting these specifications^[5]. The necessary information should exist, but data sheet sections may be named differently; they may include certain information in different groupings, and terminology may be slightly different.

Specifications may be provided in mixed combinations of minimum, typical/nominal, and maximum. When a minimum or maximum limit is not specified, it is understood to be self-evident or subject to a physical limitation that is beyond the scope of the device. Using Fairchild's 74LS00 as an example, no minimum output current is specified, because the physical minimum is very near zero. The actual output current is determined by the load that is being driven, assuming that the load draws no more than the specified maximum. Other specifications are shown under certain operating conditions. A well written data sheet provides guaranteed specifications under worst-case conditions. Here, the logic 1 output voltage (V_{OH}) is specified as a minimum of 2.5 V under conditions of minimum supply voltage (V_{CC}), maximum output current (I_{OH}), and maximum logic-low input voltage (V_{IL}). These are worst-case conditions. When V_{CC} decreases, so will V_{OH} . When I_{OH} increases, it places a greater load on the output, dragging it down to its lowest level.

Timing specifications may also be incomplete. Manufacturers do not always guarantee minimum or maximum parameters, depending on the specific type of device and the particular specification. As with DC voltages, worst-case parameters should always be specified. When a minimum or maximum delay is not specified, it is generally because that parameter is of secondary importance, and the manufacturer was unable to control its process to a sufficient level of detail to guarantee that value. In many situations where incomplete specifications are given, there are acceptable reasons for doing so, and the lack of information does not hurt the quality of the design.

Typical timing numbers are not useful in many circumstances, because they do not represent a limit of the device's operation. A thorough design must take into account the best and worst performance of each IC in the circuit so that one can guarantee that the circuit will function under all conditions. Therefore, worst-case timing parameters are usually the most important to consider first, because they are the dominant limit of a digital system's performance in most cases. In more advanced digital systems, minimum

parameters can become equally as important because of the need to meet hold time and thereby ensure that a signal does not disappear too quickly before the driven IC can properly sense the signal's logic level.

Output timing specifications are often specified with an assumed set of loading conditions, because the current drawn by the load has an impact on the output driver's ability to establish a valid logic level. A small load will enable the IC to switch its output faster, because less current is demanded of the output. A heavier load has the opposite effect, because it draws more current, which places a greater strain on the output driver.

2. 1. 2 Specialized English Words

IC	集成电路	MOS process	金属氧化物半导体
datasheet	数据手册		制程
terminology	术语学	heats up	加热
specifications	规范	threshold	阈值
absolute maximum ratings	最大绝对额	keep in mind	紧记
定值		worst-case	最坏的
parameters	参数	timing	时序

2. 1. 3 Notes

[1] Once the basic data sheet terminology and organization is understood, it is relatively easy to figure out other data sheets even when their exact terminology changes. “it” 在这里指的是 “to figure out other data sheets”。“once”引导的是条件状语从句。全句可译为“一旦掌握了基本数据手册内的术语和结构,即使以后其中一些的特殊术语改换名称,我们也很容易理解数据手册的内容。”

[2] Manufacturers specify a storage temperature range within which the semiconductor structures will not break down. 句中“within which”引起定语从句,“range”为其先行词。全句可译为“生产商通常会注明在保证半导体器件不受破坏的情况下的储藏温度范围。”

[3] Other parameters establish the safe operating limits for input signals as well as the applied voltage thresholds that represent logic 0 and 1 states. “that”引导的定语从句“represent logic 0 and 1 states”修饰“voltage thresholds”,“as well as”相当于“以及,和”。全句可译为“还有一些参数建立了输入信号的安全限制,以及代表逻辑 1 和 0 的电压阈值等。”

[4] DC parameters specify the voltages and currents that the IC will present to other circuitry to which it is connected. 由“that”引导的定语从句,修饰“voltages and currents”,从句中由“to which”又引导了一个定语从句,修饰“other circuitry”,“it”指“the IC”。全句可译为“器件的直流参数主要描述器件和其他电路连接时所表现出的电

压和电流特性。”

[5] Keep in mind that each manufacturer has a somewhat different style of presenting these specifications. “Keep in mind …”为祈使句,省略了主语“We (or you) must”。全句可译为“这里我们需要注意的是每一个生产厂商的数据手册格式多少会有一些不同。”

2.1.4 Reference Translation

半导体生产厂商对他们的每种产品都印制了相应的数据手册。不管它们是何种类型的设备系列,所有的逻辑集成电路数据手册的格式都是一样的。一旦掌握了基本数据手册内的术语和结构,即使以后其中一些的特殊术语改换名称,我们也很容易理解数据手册的内容。我们以 74LS00 数据手册为例,现摘取其中一页,如图 2.1.1 所示。

数字集成电路的数据手册至少包括两个部分:功能描述和电气说明。功能描述中的内容包括引脚配置及说明,以及该器件的详细操作描述。像 74LS00 这样一个简单的集成电路,它的功能描述非常简单,这是因为我们实在不必要对简单的与非操作进行太多的说明。但是像微处理器等复杂的集成电路,它们的功能描述则需要有几十页甚至上百张的说明,而且还要被分成许多章节。此外,一些数据手册还需要附加一些章节用于说明器件封装的物理尺寸和热学特性。大多数型号的数字集成电路的数据手册都是相同的,内容大致可分为以下四类。

绝对最大额定值。正如术语所表示的,这些参数表明了在没有永久性损伤存在的情况下集成电路器件所能承受的绝对极限。生产厂商通常对这些参数进行这样的注释:集成电路从来都不允许在这些极限条件下运行。这些额定参数非常有用,它们说明了器件的储藏温度范围,并表示了物理芯片设计和生产的品质。生产商通常会注明在保证半导体器件不受破坏的情况下的储藏温度范围。仍以 74LS00 为例,它的储藏温度范围是 -65°C 到 $+150^{\circ}\text{C}$ 。此外,对最大额定电压也进行了描述。例如 74LS00 的最大电压为 $+7\text{V}$,这说明了该器件可以承受的电压最大可达到 $+7\text{V}$ 而不受到损坏。

推荐工作条件。这些参数表明了集成电路的正常工作电压和工作温度范围。在规定的工作范围下工作,集成电路能够保证实现生产厂商所提供的功能。其中两个最重要的规定是供电电源(通常根据双极工艺还是 MOS 工艺的不同分别标志为 V_{CC} 或 V_{DD})和运行温度。一个集成电路可能有多种电源规格,这是因为一个集成电路实际上可以同时在这几种不同的电压下进行工作。每一种供应电源为芯片的不同的部分供电。当生产厂商对供应电源进行说明时,包括有一定的公差,通常是 $\pm 5\%$ 或 $\pm 10\%$ 。例如许多 5V 的逻辑 IC 只有供应电源在 $+4.75\text{V}$ 到 $+5.25\text{V}$ 之间时才能正常运行。此外,运行过程时的工作温度也是非常重要的,因为它会直接影响到器件的运行速度。当器件的温度升高时,它的运行速度就会减慢。只有当它冷却时,它的运行速度才会加快。超出了推荐的工作温度范围器件将不能正常地工作,甚至有可能由于温度过高而严重危及到器件本身。这里有四种常见的集成电路温度范围:商业应用(0°C 至 70°C)、工业应用(-40°C 至 85°C)、汽车应用(-40°C 至 125°C)以及军事应用(-55°C 至 125°C)。一般情况下

生产厂商很难生产出一个可以在温度范围外正常运行的集成电路。也正因为如此,对工作温度的要求越高,集成电路的价格也就越高。

还有一些参数建立了输入信号的安全限制,以及代表逻辑 1 和 0 的电压阈值等。这类参数通常有最大输入和最小输入两种说明,或者采用绝对电压的描述形式,或者采用相对于器件电源的相对电压形式。可以肯定的是,超出这些电压范围就会损坏器件。逻辑阈值电压的限定能够确定逻辑输入电压的大小,使器件按正常功能运行,而不会出现把逻辑 1 视为逻辑 0 的现象。这里还有对输出驱动能力的限制。我们应该熟悉这些说明,以保证器件的输出负载不能超出额定,否则将对芯片产生永久性的破坏或者不能达到规定的功能要求。

直流电学特性。器件的直流参数主要描述器件和其他电路连接时所表现出的电压和电流特性,尽管推荐工作条件给出了芯片能正常工作的环境要求,直流电学特性给出了芯片自身建立的小环境。在输出电压说明中定义了逻辑 1 和 0 的阈值电压,从而保证在所有合格的运行条件下器件能够正常地工作。这些说明保证了该芯片与同一系列中的其他芯片可以互换,也让工程师能明确该芯片的输出标准是否与它要驱动的其他芯片兼容。

输入电流说明定义了器件被其他电路驱动时的负载情况。不管输入的逻辑状态是什么,在相关器件的输入与驱动器件的输出之间存在一个小电流。通过对这些电流大小进行定量限制,可以确保多个集成电路之间的兼容性。例如当一个集成电路驱动其他几个集成电路时,输入电流的总数应不超过驱动器件的输出电流额定值。

交流电学特性(或开关特性)。数字集成电路的说明书中交流参数的复杂性和详细的程度是最高的,其中包括输入和输出的时序参数要求。像 74LS00 这样的集成电路是纯组合电路,时序参数只需要定义信号延迟以及信号的上升和下降时间。但是包含如触发器这样的时序元件的集成电路则需要标明数据建立时间、数据保持时间、时钟频率以及输出有效时间等。

这里我们需要注意的是每一个生产厂商的数据手册格式多少会有一些不同。必要信息不可缺少,但数据手册的各部分的命名可能不尽相同,信息分组也会各异,术语也会有所差别。

技术规格通常使用最小值、典型值和最大值的三种组合。当没有明确地说明最小值和最大值时,则指的是一些默认值或者超出器件能力的实际界线。仍以 74LS00 为例,它没有标明最小输出电流,这是因为实际的最小电流接近于零。如果负载没有超出数据手册中额定的电流最大值,则实际输出电流由所驱动的负载来决定。下面我们要介绍的是在一些其他工作条件下的技术规格。

好的数据手册中还会给出最恶劣的工作条件。例如在最低电源电压 V_{CC} 、最大输出电流 I_{OH} 和最大逻辑低电平电压 V_{IL} 情况下,逻辑 1 的输出电压 V_{OH} 的最小值为 2.5 V。这些就是最恶劣的条件。当 V_{CC} 下降时, V_{OH} 也会下降。当 I_{OH} 上升时,增加了输出负载,又将 V_{OH} 拉向最低值。

时序说明也可能是不完整的。对于特殊类型的设备和特定的规格,生产厂商并不能一直保证最大和最小参数。对直流电压来说,最坏情况下的参数说明是必不可少的。如

果没有特别指明最小或最大延迟,一般情况下是由于这些参数不重要。此外生产商也无法精确控制其加工工艺确保达到该参数指标。在很多情况下不完整的说明也是可以接受的,这主要是因为缺少某些信息并不会妨碍器件的设计品质。

在很多情况下,时序参数典型值并没有多大的用处,因为典型值并非是器件运行的一种限制条件。一个充分的设计应该充分考虑最坏和最好两种情况,从而保证器件在所有的条件下均能正常工作。通常最坏情况下的时序参数是最重要的,必须首先考虑。主要原因是因为在大多数情况下最坏情况限定了整个数字系统的性能。此外,在一些更加先进的数字系统设计中最小参数也会变得同样重要。这是为了满足数据保持时间的需要,从而确保在下一级电路完成信号采样前信号不会过快地消失。

由于负载上的电流对输出驱动建立有效的逻辑电平的能力有很大影响,因此输出的时序说明通常是在一系列假定的负载情况下进行描述的。轻负载可以使集成电路的输出信号转换速度更快,此时输出驱动只需要少量的电流。然而,重负载的影响相反,因为它吸纳了更多的电流,给输出驱动造成巨大的负担。

2.1.5 Reading Materials

The Integrated Circuit

Digital logic and electronic circuits derive their functionality from electronic switches called transistors. Roughly speaking, the transistor can be likened to an electronically controlled valve whereby energy applied to one connection of the valve enables energy to flow between two other connections. By combining multiple transistors, digital logic building blocks such as AND gates and flip-flops are formed. Transistors, in turn, are made from semiconductors. Consult a periodic table of elements in a college chemistry textbook, and you will locate semiconductors(半导体) as a group of elements separating the metals and nonmetals. They are called semiconductors because of their ability to behave as both metals and nonmetals. A semiconductor can be made to conduct electricity like a metal or to insulate as a nonmetal does. These differing electrical properties can be accurately controlled by mixing the semiconductor with small amounts of other elements. This mixing is called doping. A semiconductor can be doped to contain more electrons (N-type) or fewer electrons (P-type). Examples of commonly used semiconductors are silicon and germanium. Phosphorous and boron are two elements that are used to dope N-type and P-type silicon, respectively.

A transistor is constructed by creating a sandwich of differently doped semiconductor layers. The two most common types of transistors, the bipolar junction transistor (BJT) and the field-effect transistor (FET,场效应管) are schematically illustrated in Figure 2.1.2. This figure shows both the silicon structures of these

elements and their graphical symbolic representation as would be seen in a circuit diagram. The BJT shown is an NPN transistor, because it is composed of a sandwich of N-P-N doped silicon. When a small current is injected into the base terminal, a larger current is enabled to flow from the collector to the emitter. The FET shown is an N-channel FET; it is composed of two N-type regions separated by a P-type substrate. When a voltage is applied to the insulated gate terminal, a current is enabled to flow from the drain to the source. It is called N-channel, because the gate voltage induces an N-channel within the substrate, enabling current to flow between the N-regions.

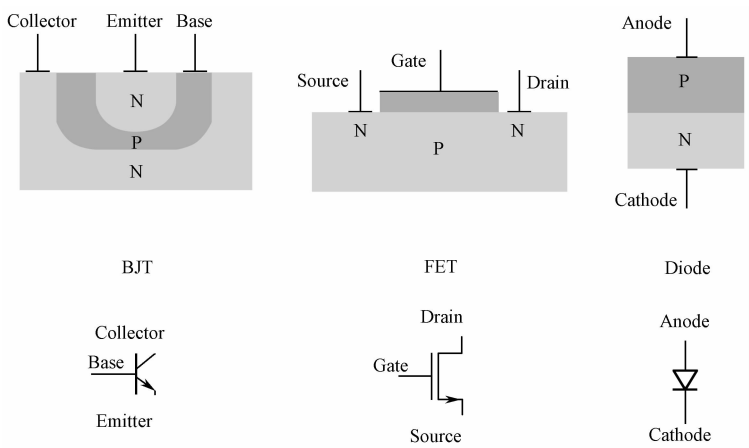


Figure 2. 1. 2 Two most common types of transistors.

2. 2 Diodes and Transistors(I)

2. 2. 1 Text

Most of the semiconductors that a digital system employs are fabricated as part of integrated circuits. Yet there are numerous instances in which discrete semiconductors, most notably diodes and transistors, are required to complete a system^[1]. Diodes are found in power supplies, where they serve as rectifiers and voltage references. It is difficult these days not to find a light emitting diode, or LED, in one’s immediate vicinity, on some appliance or piece of electronic equipment^[2]. Discrete transistors are present in switching power supplies and in circuits wherein a digital IC must drive a heavy load. There are many other uses for diodes and transistors in analog circuit design, most notably in signal amplification. These more analog topics are not discussed here.

An ideal diode is a nonlinear circuit element that conducts current only when the

device is forward biased , i. e. , when the voltage applied across its terminals is positive. It thereby behaves as a one-way electrical valve that prevents current from flowing under conditions of reverse bias . A diode has two terminals; the anode and cathode . For the diode to be forward biased, the anode must be at a more positive voltage than the cathode. Diodes are the most basic semiconductor structures and are formed by the junctions of two semiconductor materials of slightly differing properties. In the case of a silicon diode, the anode is formed from positively doped silicon, and the cathode is formed from negatively doped silicon. Along this pn junction is where the physical phenomenon occurs that creates a diode. Figure 2. 2. 1 shows the general silicon structure of a diode and its associated symbolic representation.

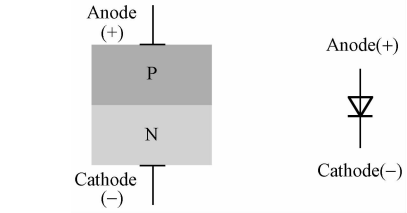


Figure 2. 2. 1 Diode structure and graphical representation.

Real diodes differ from the ideal concept in several ways. Most significantly, a real diode must be forward biased beyond a certain threshold before the device will conduct. This threshold is called the forward voltage, V_F . A diode conducts very little current below V_F , measured in micro- or nanoamps. The relationship between a diode’s current and voltage is exponential and therefore increases rapidly around V_F . Above V_F , the

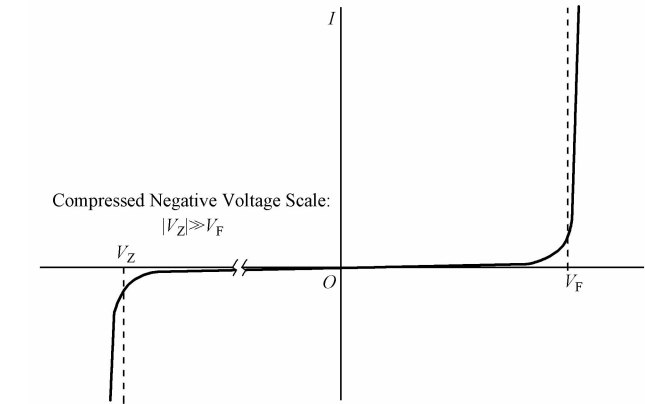


Figure 2. 2. 2 Silicon diode I - V characteristic.

diode presents very low impedance and appears almost like a short circuit, or a piece of wire. A typical silicon diode’s I - V characteristic is shown in Figure 2. 2. 2 with V_F of approximately 0.7 V. For many applications, especially those in digital systems, a diode’s I - V characteristic can be simplified to a step function from zero to infinite current at the forward voltage. Real diodes also have a reverse breakdown voltage, called the Zener voltage, V_Z , at which point they will conduct under reverse-biased conditions. Normal diodes are usually not subjected to reverse bias voltages sufficient to achieve significant conduction. Special-purpose Zener diodes are designed specifically to operate under reverse-bias conditions and are commonly used in voltage reference and regulation applications.

Saying that a diode conducts “infinite” current beyond V_F really means that its impedance becomes so low that it no longer becomes the limiting factor in a circuit^[3]. If a typical diode is connected directly to a battery such that it is forward biased, the diode will form nearly a short circuit, which will cause a large current to flow. Fairly soon afterward, the diode will likely fail due to thermal overload. Diodes are specified with maximum forward currents, I_F . Exceeding I_F causes the diode to dissipate more power than it is designed for, usually with destructive consequences.

When a diode is used in an application wherein it may be forward biased and driven with excessive current, a current-limiting resistor is inserted into the circuit to keep the diode within its specified operating limits^[4]. Diodes are useful for providing a fixed voltage reference regardless of a circuit’s operating voltage. The circuit shown in Figure 2. 2. 3 takes advantage of a diode’s relatively static forward voltage with respect to current. A loosely regulated 12 V supply may have a tolerance of ± 20 percent—a range of 9. 6 to 14. 4 V. If a resistor divider is used to generate a reference voltage, its accuracy could be no better than that of the 12 V supply. Some applications require a more accurate voltage reference with which to sense an incoming signal. A 1N4148 exhibits $V_F = 0. 7$ V at 5 mA under typical conditions. The 2. 2 k Ω resistor limits the current through the diode to approximately 5 mA when $V_{IN} = 12$ V. If the input changes by 20 percent and causes a corresponding change in the current, the diode voltage changes by a small fraction. Using a basic small signal diode in this manner is an effective scheme for many applications. If tighter tolerance is desired, more stable voltage reference diodes are available. Thermal overload is not a problem for this diode, because its power dissipation is relatively constant at $0. 7\text{ V} \times 5\text{ mA} = 3. 5\text{ mW}$.

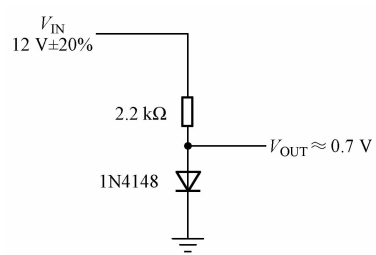


Figure 2. 2. 3 Diode-based voltage reference.

Diodes are available with a broad spectrum of characteristics. Aside from silicon diodes, there are Schottky diodes that exhibit lower forward voltages of under 0. 5 V. Lower forward voltages provide benefits for high-power applications in which heat and power dissipation are prime concerns. Reduced V_F means reduced power.

Diodes are manufactured in a variety of packages according to the amount of power that they are designed to handle. Small-signal diodes are not intended to handle much power and are available in small, surface mount packages. At the other extreme, diodes can be as large as hockey pucks for very high-power applications. Small-signal diodes are

manufactured with varying response times to changes in voltage. A diode can be used to clip a signal to prevent it from exceeding a certain absolute voltage, as shown in Figure 2.2.4. As the signal's edge rate increases, a slower diode may not respond fast enough to be effective. If a single diode's forward voltage is insufficient, multiple diodes can be placed in series to increase the clipping threshold. Some of the more common small-signal diodes used in digital circuits include the leaded 1N914 and 1N4148 devices, and their surface mount equivalents, the SOT-23 MMBD914 and MMBD4148.

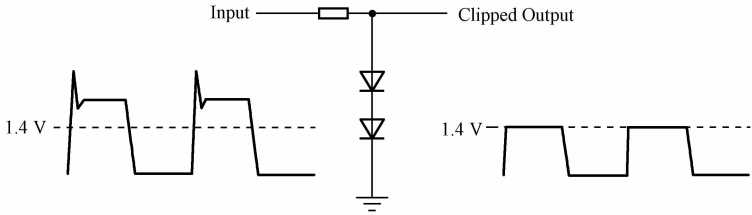


Figure 2.2.4 Clipping a signal with a diode.

A major use for diodes is in the rectification of AC signals, specifically in power supplies in which the conversion from AC to DC is required^[5]. Small-signal diodes can be used as rectifiers in nonpower or low-power applications. Larger diodes with higher power ratings are employed when constructing power supply circuits meant to provide more power. An AC power signal is a sine wave of arbitrary amplitude that is centered about 0 V. Its voltage peaks are of equal magnitude above and below 0 V. A digital circuit requires a steady DC power supply. The first step in creating a steady DC power supply is to rectify the AC input such that the negative AC sine wave excursions are blocked. Figure 2.2.5 shows a single diode performing this function. The rectified output is reduced in voltage by the diode's forward voltage. This circuit is called a half-wave rectifier, because it passes only half of the incoming power signal. Once rectified, capacitors and inductors can smooth out (lowpass filter) the rectified AC signal to create a steady DC output.

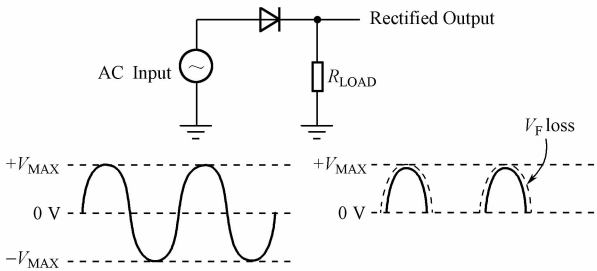


Figure 2.2.5 Half-wave rectifier circuit.

2.2.2 Specialized English Words

semiconductors	半导体	discrete	分立的, 离散的
fabricated	制作, 构成	diodes	二极管

transistors 晶体管
power supplies 电源
rectifiers 整流器
voltage references 参考电压
light emitting diode 发光二极管
switching 开关
heavy load 重负载
signal amplification 信号放大
ideal diode 理想二极管
one-way 单向
valve 电子管

reverse bias 反偏
forward biased 正偏
junctions 结
Real diodes 实际的二极管
threshold 阈值
impedance 阻抗
short circuit 短路
dissipate 消耗
current-limiting resistor 限流电阻
Thermal overload 热过载
lowpass filter 低通滤波器

2. 2. 3 Notes

[1] Yet there are numerous instances in which discrete semiconductors, most notably diodes and transistors, are required to complete a system. 在这个主从复合句中,主句为“Yet there are numerous instances”。“in which…”为定语从句,修饰“instances”。从句中“most notably diodes and transistors”为“discrete”的同位语。全句可译为“但仍有很多使用分立式半导体元件,主要是二极管和三极管,来构成系统的例子。”

[2] It is difficult these days not to find a light emitting diode, or LED, in one's immediate vicinity, on some appliance or piece of electronic equipment. 本句中采用了双重否定“It is difficult … not to…”来强调 LED 的广泛应用。全句可译为“如今在我们的身边,例如各种器械和电子设备上,到处都可以见到发光二极管或者 LED。”

[3] Saying that a diode conducts “infinite” current beyond V_F really means that its impedance becomes so low that it no longer becomes the limiting factor in a circuit. 本句结构较复杂,主语为“Saying that … beyond V_F ”,“Saying”为动名词,它的宾语为“that”引导名词从句。“means”为主语“Saying …”的谓语动词,而“that its impedance…”为“means”的宾语从句,而这个从句又有自己的结果状语“so that…”。全句可译为“我们说二极管在 V_F 以上导通的电流无穷大实际上是指它的阻抗变得如此之小以致它不再是电路中的限制因素。”

[4] When a diode is used in an application wherein it may be forward biased and driven with excessive current, a current-limiting resistor is inserted into the circuit to keep the diode within its specified operating limits. 这是一个多层的主从复合句。主句为“a current-limiting resistor…”,“When a diode is used in an application”为主句的状语从句,而从句中的“an application…”又带有一定语从句“wherein it may be…”,其中有两个并列被动语态的谓语“be forward biased”和“driven”。全句可译为“在应用中如果二极管正向偏置时的驱动电流很大,就要在电路中放置一个限流电阻以保证二极管工作在它的规范值之内。”

[5] A major use for diodes is in the rectification of AC signals, specifically in power supplies in which the conversion from AC to DC is required. 在这个典型主从复合句中, “in which” 引起定语从句, 修饰“power supplies”。全句可译为“二极管的一个主要应用是 AC 信号整流, 尤其是在某些需要将 AC 电转化为 DC 电的电源中。”

2.2.4 Reference Translation

二极管和三极管(I)

数字系统中使用的绝大多数半导体元件都被制作成集成电路的一部分。但仍有很多使用分立式半导体元件, 主要是二极管和三极管, 来构成系统的例子。在电源中我们常可以见到二极管, 它们的用途是作为整流器和基准电压。如今在我们的身边, 例如各种器械和电子设备上, 到处都可以见到发光二极管或者 LED。分立式三极管常用于开关电源或者需要驱动很大负载的数字 IC 电路中。在模拟电路设计中也常常用到二极管和三极管, 最主要是用来放大信号。关于模拟电路方面的深入内容, 这里就不讨论了。

理想的二极管是一个非线性元件, 只有当它处于正向偏置时(例如在它两端施加正向偏压)才有导通电流。因此它就像一个单向阀, 在反向偏置时阻止电流的流动。一个二极管有两个端: 正极和负极。如果要使二极管处于正向偏置, 正极的电压就必须高于负极的电压。二极管是一种很基本的半导体结构, 它是把两种特性稍有不同的半导体材料结合在一起所构成的。对于一个硅二极管, 正极由正掺杂的硅构成, 负极则由负掺杂的硅构成。二极管就是利用在 PN 结上产生的特殊物理现象而工作的。图 2.2.1 展示了普通硅二极管的结构和它的符号表示法。

实际的二极管和理想概念有许多不同, 其中最重要的是, 一个实际的二极管只有在正向偏置超过一个阈值后才会导通。这个阈值称为正偏开启电压(V_F)。在小于 V_F 时二极管中的电流非常小, 处于微安或者纳安量级。二极管的电流和电压之间是指数关系, 因此在 V_F 点以后电流增加得非常快。高于 V_F 后, 二极管表现出很小的阻抗, 几乎就如同短路或者一根导线一样。典型的硅二极管的 $I-V$ 特性如图 2.2.2 所示, V_F 约等于 0.7 V。在许多应用中, 特别是数字系统中, 二极管的 $I-V$ 特性可以被简化为分段函数, 在 V_F 点发生了从零电流到无穷大电流的突变。实际中的二极管还存在一个反向击穿电压, 称为齐纳电压(V_Z), 在这一点出现反向偏置而导通。一般的二极管通常不会工作在很大的反向偏压下。但特殊设计的齐纳二极管则专门工作在反向偏压下, 通常用于基准电源和稳压中。

我们说二极管在 V_F 以上导通的电流无穷大实际上是指它的阻抗变得如此之小, 以致它不再成为电路中的限制因素。如果一个典型的二极管直接被连接到电池形成正向偏置, 它就像短路了一样, 会产生很大的电流。这个二极管很快就会因为过热而烧毁。每个二极管都有一个最大正向电流(I_F)指标, 超过 I_F 会导致二极管消耗的功率大于设计值,

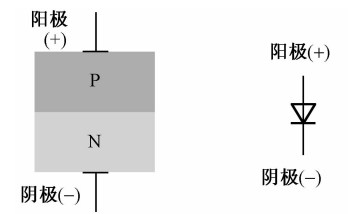


图 2.2.1 二极管结构和图形符号

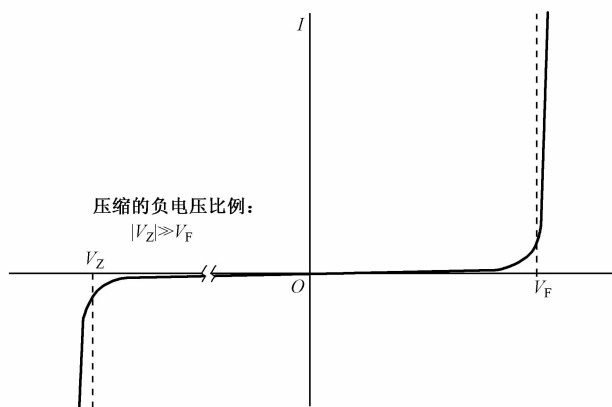


图 2.2.2 硅二极管的伏-安特性

生基准电压,它的准确度不会高于 12 V 电源。某些应用需要一个更准确的基准电压来感应输入信号。一个 1N4148 在 5 mA 的典型情况下, $V_F = 0.7$ V。一个 $2.2\text{ k}\Omega$ 的电阻在 $V_{IN} = 12$ V 的情况下将流过二极管的电流限制为约等于 5 mA。如果输入电源电压在 20% 的范围内变化导致电流变化,二极管的电压值也只会产生很小的变化。在很多应用场合中,照这样使用一个基本小信号二极管都是一个很有效的方案。如果容差要求更严,就可以使用更稳定的基准电压二极管。对于这个二极管,不会出现过热的问题,因为它的功耗基本是一个恒定值: $0.7\text{ V} \times 5\text{ mA} = 3.5\text{ mW}$ 。

有各种各样特性的二极管可供应用。除了硅二极管,肖特基二极管具有低于 0.5 V 的正偏导通电压。较低的正向导通电压对主要考虑热量和功耗问题的大功率应用十分有益,因为降低 V_F 意味着降低功耗。根据不同的功率应用场合,二极管具有不同的封装形式。小信号二极管由于应用场合的功率不大,所以采用很小的表面贴片封装。相反,在大功率应用场合中二极管可以像冰球一样大。小信号二极管的电压变化响应时间有大有小。二极管可以用来对一个信号限幅,使之不超过某一个特定电压值,如图 2.2.4 所示。当这个信号的变化速率加快时,反应速度较慢的二极管可能跟不上这种变化而效能下降。如果单个二极管的正偏导通电压不够的话,可以通过串联多个二极管来达到所要求的限幅电压。在数字电路中常用的小信号二极管包括有引线的 1N914 和 1N4148 系列,以及和它们等效的表面贴装形式的 SOT-23 MMBD914 和 MMBD4148 系列。

二极管的一个主要应用是 AC 信号整流,尤其是在某些需要将 AC 电转化为 DC 电的电源中。小信号二极管可以在无功率或者低功率应用中作为整流器。具有较大功率的二极管可以用来构成大功率电源电路。AC 电源信号是一个平均值为 0 V 的正弦波,无

通常的结果是二极管烧毁。

在应用中如果二极管正向偏置时的驱动电流很大,就要在电路中放置一个限流电阻以保证二极管工作在它的规范值之内。二极管可以提供一个稳定的基准电压而无论电路的实际电压是多大。图 2.2.3 所示的电路就利用了二极管的正偏导通电压相对电流保持不变的特性。一个只经过简单稳压的 12 V 电源可能有 $\pm 20\%$ 的容差,电压范围从 9.6 V 到 14.4 V。如果用一个电阻分压器来产

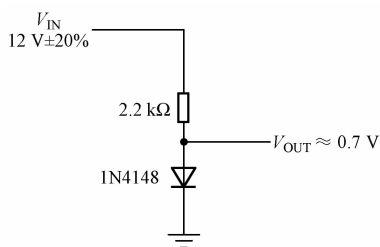


图 2.2.3 基于二极管的参考电压源

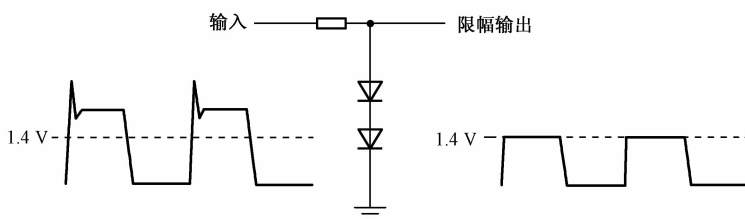


图 2.2.4 二极管信号限幅电路

论幅值是多少。它的正负电压峰值是相等的。数字电路需要一个稳定的 DC 电源供应。建立稳定的 DC 电源的第一步是对 AC 输入进行整流,去除负半周的 AC 正弦信号。图 2.2.5 中所示的单个二极管可以实现这个功能。整流后的输出电压由于二极管正偏导通电压的损耗而略有减小。这个电路称为半波整流器,因为它只让输入电源信号的一半通过。然后,可以用电容和电感来平滑(低通滤波)整流后的 AC 信号,从而得到一个稳定的 DC 输出。

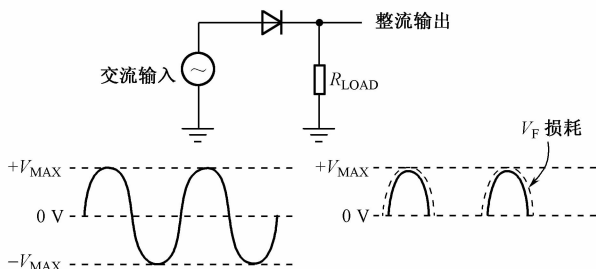


图 2.2.5 半波整流电路

2.2.5 Reading Materials

IC Packaging(封装)

When the wafer(晶圆) has completed its final process step, it is tested and then sliced up to separate the individual dice. Dice that fail the initial testing are quickly discarded. Those that pass inspection are readied for packaging. A package is necessary for several reasons, including protection of the die and the creation of electromechanical connections with other circuitry. ICs are almost always mounted onto a circuit board, and it is usually difficult to mount unpackaged ICs directly to the board. However, there are special situations in which ICs are not packaged and are directly attached to the board. These cases are often at opposite ends of the technological spectrum. At the low end of technology, ICs can be several process generations behind the current state of the art. Therefore, the relative complexity of mounting them to a circuit board may not be as great. The savings of direct mounting are in space and cost. A common quartz wristwatch benefits from direct mounting, because the small confines of a watch match very well with the space savings achieved by not requiring a package for the IC. These watch ICs use mature semiconductor process technologies. At the high end of

technology, some favorable electrical and thermal characteristics can be achieved by eliminating as much intermediate bulk as possible between individual ICs and supporting circuitry. However, the technical difficulties of direct-mounting a leading-edge IC can be challenging and greatly increase costs. Therefore, direct-mounting of all but very low-end electronics is relatively rare.

IC packaging technology has evolved dramatically from the early days, yet many mature package types still exist and are in widespread use. Plastic and ceramic are the two most common materials used in an IC package. They surround the die and its lead frame. The lead frame is a structure of metal wires that fan out from the die and extend to the package exterior as pins for connection to a circuit board. Plastic packages are generally lower in cost as compared to ceramics, but they have poorer thermal performance. Thermal characteristics are important for ICs that handle large currents and dissipate large quantities of heat. To prevent the IC from overheating, the heat must be conducted and radiated away as efficiently as possible. Ceramic material conducts heat far better than plastic.

2.3 Diodes and Transistors(II)

2.3.1 Text

The single-diode half-wave rectifier does the job, but does not take advantage of the negative portion of the AC input. Four diodes can be assembled into a full-wave bridge rectifier that passes the positive portion of the sine wave and inverts the negative portion relative to the DC ground^[1]. This circuit is shown in Figure 2.3.1. The bridge rectifier works by providing a current conduction path through the resistor to ground regardless of the polarity of the incoming AC signal. When the AC input is positive with respect to the polarity markings shown in the diagram, diodes D1 and D3 are forward biased, conducting current from D1 through the resistor, then through D3 to the negative AC input wire. When the AC input is negative during the next half of the sinusoid, D2 and D4 are forward biased and allow current to flow in the same direction through the resistor. The result is that a positive voltage is always developed across the load with respect to ground. Note that, because of the two diodes in series with the load, the rectified output voltage is reduced by twice the diodes' forward voltage.

Power rectifier circuits are generally found in systems wherein a high-voltage input (e. g. , 120 V AC) must be converted into a low-voltage output such as +5 V DC to power a digital logic circuit. Transformers are used in conjunction with bridge rectifiers to step down the high-voltage AC input to a more appropriate intermediate level that is much closer to the final voltage level required by the system^[2]. A power filter circuit

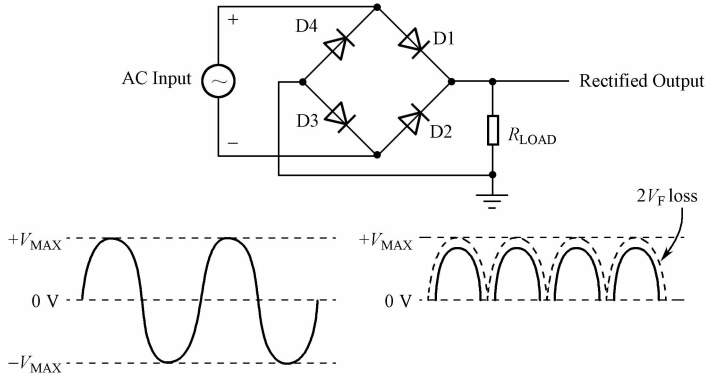


Figure 2.3.1 Bridge rectifier circuit.

can then be used to smooth the heavily rippled rectified signal into a more stable DC input. Finally, a voltage regulator performs the final adjustments to convert the intermediate voltage into a more accurate digital supply voltage. This common AC-to-DC power supply configuration is illustrated in Figure 2.3.2.

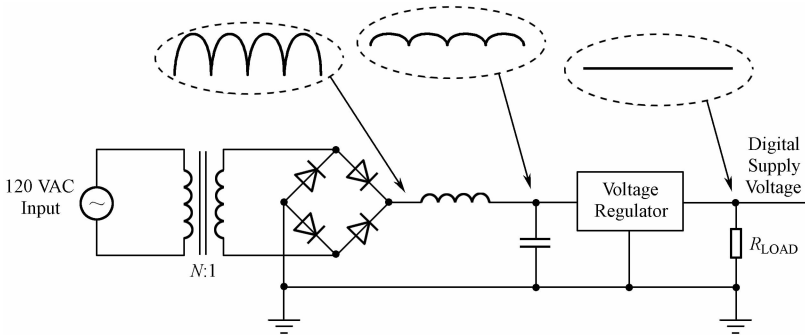


Figure 2.3.2 AC-to-DC Power Supply.

Another power application of diodes is in combining multiple power supplies to feed a single component or group of components while ensuring that the failure or disappearance of one supply does not cause that component to lose power and cease operation. This concept relies on the fact that a standard diode will not conduct under normal reverse-bias conditions. As shown in Figure 2.3.3, each power supply is isolated by a diode whose cathodes form a common voltage supply node for a circuit. Under normal operating conditions, each diode is forward biased, because the respective power supplies are functioning. When one supply fails, its associated diode becomes reverse biased, thereby preventing the failing supply from pulling power from the functioning supply and causing the system to fail. These diodes are often called OR-ing

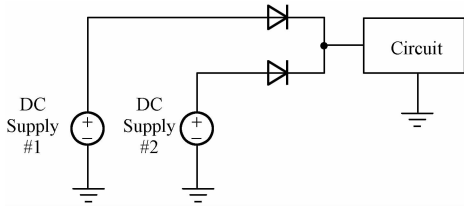


Figure 2. 3. 3 Power Supply OR-ing diode.

diodes, because they perform a logical OR function on the power supplies.

Diode OR-ing circuits are also seen in battery-backup applications in which it is desired to keep a low-power static RAM chip powered by a battery when the main power supply is turned off^[3]. A typical scenario is a higher-voltage operating supply (e.g. ,

+5 V) and a lower-voltage data-retention battery (e.g. , +3 V) are each connected to an SRAM via independent OR-ing diodes. Under normal operation, the operating supply forces the battery's diode into reverse bias, preventing the battery from supplying power to the SRAM, thereby extending the battery's life. When power is turned off, the battery's diode becomes forward biased and maintains power to the SRAM so that its data are not lost^[4]. Schemes like this are commonly employed in certain PCs and other platforms that benefit from storing configuration information in nonvolatile SRAM.

Transistors are silicon switches that enable a weak signal to control a much larger current flow, which is the process of amplification; magnifying the amplitude of a signal^[5]. Bipolar junction transistors (BJTs) are a basic type of transistor and are formed by two back-to-back pn junctions. Figure 2. 3. 4 shows the general BJT structures and their associated symbolic representations. The BJT consists of three layers, or regions, of silicon in either of two configurations: NPN and PNP. The middle region is called the base , and the two outer regions are separately referred to as the collector and emitter . As will soon be shown, the base-emitter junction is what enables control of a potentially large current flow between the collector and emitter with a very small base-emitter current. A BJT's construction is more than simply placing two diodes back to back. The base region is extremely thin to enable conduction between the collector and emitter, and the collector and emitter are sized differently according to the fabrication process.

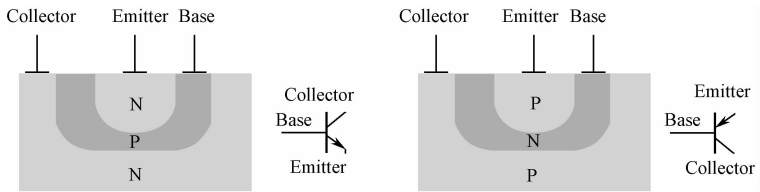
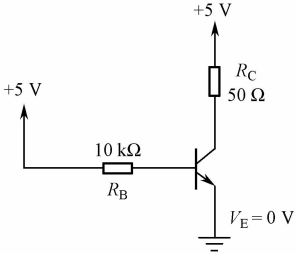


Figure 2. 3. 4 NPN and PNP BJT structures and graphical representations.

Currents in an NPN transistor flow from the base to the emitter and from the collector to the emitter. The relationship between these currents is defined by a

proportionality constant called beta (β , also known as h_{FE}): $I_C = \beta I_B$. Beta is specific to each type of transistor and is characterized by the manufacturer in data sheets. Typical values for beta are from 100 to less than 1000. The beta current relationship provides a quick view of how a small base current can control a much larger collector current. A higher beta indicates greater potential for signal amplification. Because the base-emitter junction is essentially a diode, it must be sufficiently forward biased for the transistor to conduct current ($V_{BE} = 0.7\text{ V}$ under typical conditions). A PNP transistor functions similarly, although the polarities of currents and voltages are reversed.

When a transistor circuit is designed, care must be taken not to overdrive the base-emitter junction. Like any other diode, it presents very low impedance beyond its forward voltage. Without some type of current limiting, the transistor will overheat and become damaged. Transistors are biased using resistors placed at two or three of its terminals to establish suitable operating voltages. Figure 2.3.5 shows a common NPN configuration at DC with a current limiting resistor, R_B , at the base and a voltage-



dropping resistor, R_C , at the collector. The emitter is grounded, establishing the base voltage, V_B , at 0.7 V. R_B sets the current flowing into the base and thereby controls the collector voltage, V_C . As R_B increases, V_C increases, because less current is pulled through the collector, reducing the voltage drop across R_C . In this example, the base current, I_B , is $(5\text{ V} - V_B) / R_B = 0.43\text{ mA}$. Assuming a beta of 100, the collector current, I_C , is 43 mA, and $V_C = 5\text{ V} - I_C R_C = 2.85\text{ V}$.

The transistor is limited in how much current it can drive by both its physical characteristics and the manner in which it is biased. Physically speaking, a transistor will have a specified maximum power dissipation beyond which it will overheat and eventually become damaged. In this circuit, the transistor's power dissipation is $V_{CE} I_C + V_{BE} I_B$, although the dominant term is between the collector and emitter, where the great majority of the current flow exists. Using this small simplification, the transistor is dissipating approximately $2.85\text{ V} \times 43\text{ mA} = 123\text{ mW}$.

Assuming that a transistor is not operated beyond its physical limitations, the bias configuration places an upper limit on how much current flows into the collector. A BJT has three modes of operation; cutoff, active, and saturation. In cutoff, the transistor is not conducting, because the base-emitter junction is either reverse biased or insufficiently forward biased. The collector is at its maximum voltage, and the base-collector junction is reverse biased, because no current is flowing to create a voltage

drop through R_C or its equivalent. When the base-emitter junction is forward biased, the transistor conducts current and V_C begins to drop. The transistor is in active mode. As long as the base-collector junction remains reverse biased, increasing base current will cause a corresponding increase in collector current, and the transistor remains in active mode^[6]. If I_B is increased to the point at which the base-collector junction is forward biased (increased I_C causes V_C to approach V_E), the transistor enters saturation and no longer can draw more current through the collector. Saturation does not damage the transistor, but it results in a nonlinear relationship between I_B and I_C , nullifying the effect of beta. If R_C is increased or decreased, saturation occurs at lower or higher I_C , respectively. Amplifier circuits must avoid saturation to function properly because of the resulting nonlinearity. When used in a purely digital context, however, transistors can be driven from cutoff to saturation as long as the power dissipation specifications are obeyed.

2.3.2 Specialized English Words

negative portion	负半周	proportionality constant	比例常数
full-wave bridge rectifier	全波桥式整流	base	基极
transformers	变压器	collector	集电极
in conjunction with	与……连接	emitter	发射极
bipolar junction	双极结		

2.3.3 Notes

[1] Four diodes can be assembled into a full-wave bridge rectifier that passes the positive portion of the sine wave and inverts the negative portion relative to the DC ground. 这是一个主从复合句,“that”引起定语从句,修饰“a full-wave bridge rectifier”,从句中有两个并列的谓语动词“passes”和“inverts”。全句可译为“四个二极管可以组合成一个全波桥式整流器,它能让正半周的正弦波通过,并把负半周的正弦波反转到正半周。”

[2] Transformers are used in conjunction with bridge rectifiers to step down the high-voltage AC input to a more appropriate intermediate level that is much closer to the final voltage level required by the system. 在这个主从复合句中,“that…”为定语从句,修饰“a more appropriate intermediate level”,从句中“required by the system”为过去分词短语做后置定语,修饰“the final voltage level”。全句可译为“将变压器和桥式整流器连接在一起,把高压 AC 输入降低到与系统所需要的电压值很接近的一个中间值。”

[3] Diode OR-ing circuits are also seen in battery-backup applications in which it is desired to keep a low-power static RAM chip powered by a battery when the main power supply is turned off. 本句结构上比较复杂,可以分为三个层次,即主句带有一定语从句,而从句又有自己的状语从句。“in which”以下修饰“battery-backup applications”。“when”引起状语从句,做从句中“keep”的时间或条件状语。全句可译为“二极管电路还

被用在备用电池中,以保证主要电源被关闭后电池仍可以继续向一个低功耗的静态 RAM 芯片供电。”

[4] When power is turned off, the battery’s diode becomes forward biased and maintains power to the SRAM so that its data are not lost. 本句为主从复合句,主句有两个并列谓语动词“becomes”和“maintains”。“When”引起的是条件从句,而“so that”引起的是表示结果的从句,全句可译为“当工作电源关闭后,连接电池的二极管变成正偏,替代工作电源给 SRAM 提供电能,保证了数据不会丢失。”

[5] Transistors are silicon switches that enable a weak signal to control a much larger current flow, which is the process of amplification; magnifying the amplitude of a signal. 句中“Transistors are silicon switches”为主句,“that enable a weak signal to control a much larger current flow”为“silicon switches”的定语从句。句末“which is the process of amplification; magnifying the amplitude of a signal”则为非限制性定语从句,修饰前面提到的整个从句。全句可译为“三极管是一个用弱信号来控制另一个大得多的电流的硅开关,这个过程实际上就是将信号的幅度放大的过程。”

[6] As long as the base-collector junction remains reverse biased, increasing base current will cause a corresponding increase in collector current, and the transistor remains in active mode. 本复合句有两个并列主句,共同有一个状语“As long as”。最后一个“as”为连接词,后接一从句。全句可译为“只要基极-集电极结还处于反偏状态,上升的基极电流就会相应地引起集电极电流的上升,此时的三极管仍处于放大状态。”

2.3.4 Reference Translation

二极管和三极管(II)

单个二极管的半波整流器虽然能够完成这项工作,但没有使负半周的 AC 输入得到利用。四个二极管可以组合成一个全波桥式整流器,它能让正半周的正弦波通过,并把负半周的正弦波反转到正半周。这个电路如图 2.3.1 所示。无论输入 AC 信号的极性如何,桥式整流器都能提供一条从电阻到地的电流通路。当 AC 输入如图 2.3.1 中所标为

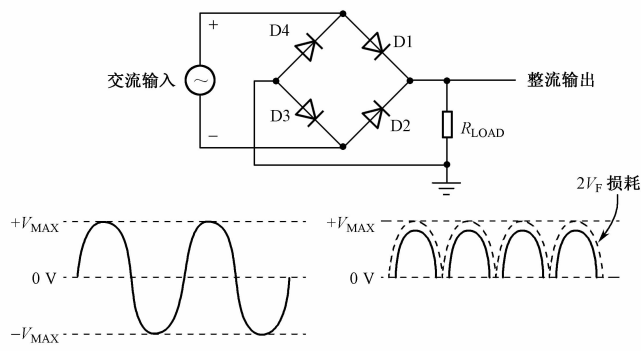


图 2.3.1 桥式整流电路

正向时,二极管 D1 和 D3 正偏,电流从 D1 流经电阻然后再流过 D3 回到 AC 输入的负极。当 AC 输入在后半个正弦周期变为反向时, D2 和 D4 正偏,使得电流仍按原方向流过电阻。最后的结果是负载到地之间总是得到正电压。但必须注意,因为两个二极管与负载之间是串联的关系,整流后的输出电压减小了两倍的正偏导通电压值。

在需要将高压输入(例如 120 V AC)转化为驱动数字电路的低压输出(例如 +5 V DC)的系统中都会见到电源整流电路。将变压器和桥式整流器连接在一起,把高压 AC 输入降低到与系统所需要的电压值很接近的一个中间值,然后通过电源滤波电路将整流后具有严重波纹的信号平滑化,得到稳定的 DC 输入。最后用电压调整器进行最终的调整,将这个电压中间值转化为更精确的数字电路电压。这个常用的将 AC 电源转化为 DC 电源的配置如图 2.3.2 所示。

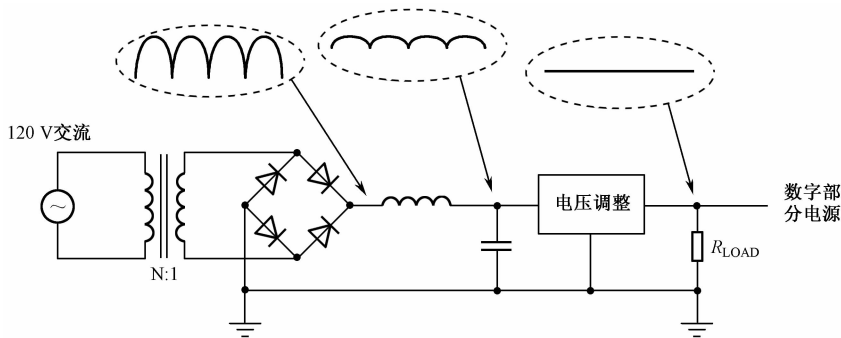


图 2.3.2 交流到直流转换电路

二极管的另一个电源应用是将多个电源同时向一个独立元件或者一组元件供电,以保证即使其中一个电源失效或者发生故障时元件不会因为失去电源供应而停止工作。这

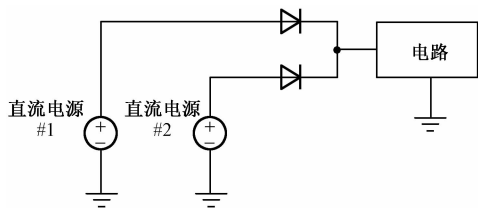


图 2.3.3 电源中的“或”二极管

有赖于标准的二极管在正常的反向偏置下不导通这一特性。如图 2.3.3 所示,每一个电源都由一个二极管隔开,二极管的负极接在一起,成为向电路供电的公共点。在正常情况下,每一个二极管都是正向偏置的,因为它们对应的电源都在正常工作。当其中一个电源失效时,和它相连的二极管就变成了反向偏置,阻止了电能从正常工作的电源流向失效的电源,因此也就阻止了整个系统的失效。这些二极管通常称为“或”二极管,因为它们电源供应中起逻辑“或”的作用。

二极管电路还被用在备用电池中,以保证主要电源被关闭后电池仍可以继续向一个低功耗的静态 RAM 芯片供电。典型的应用是在 SRAM 上通过两个独立的“或”二极管分别连接一个高电压工作电源(例如 +5 V)和一个用于保持数据的低电压电池(例如 +3 V)。在正常情况下,工作电源使连接电池的二极管反偏,阻止了电池给 SRAM 供应电能,因此延长了电池的寿命。当工作电源关闭后,连接电池的二极管变成正偏,替代工

作电源给 SRAM 提供电能,保证了数据不会丢失。像这样的方案常常用在某些 PC 和其他平台上,用来在非易失性 SRAM 中存储信息。

三极管是一个用弱信号来控制另一个大得多的电流的硅开关,这个过程实际上就是将信号的幅度放大的过程。双极型三极管(BJT)是三极管的一种基本类型,它由两个背对背的 PN 结构成。图 2.3.4 展示了一般 BJT 的结构和它的表示符号。BJT 三极管是由三个硅层或者说硅区构成,分为两种类型: NPN 和 PNP。中间的区域称为基极,两端的区域分别称为集电极和发射极。下面将要说明,基极-发射极结是如何通过一个很小的基极发射极电流来控制集电极和发射极之间的大电流的。三极管的结构并非是简单地将两个二极管背对背地放置在一起。三极管的基区非常薄,可以使集电区和发射区导通,而集电区和发射区则由于制造工序的不同在尺寸上也不同。

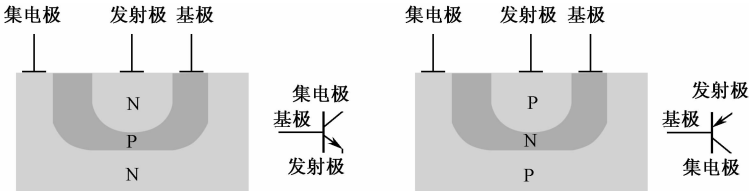


图 2.3.4 NPN 和 PNP 双极型三极管结构与图形符号

在 NPN 三极管中,电流从基极和集电极分别流向发射极。这两个电流的关系由一个称为 β (也称为 h_{FE}) 的比例常数来确定: $I_C = \beta I_B$ 。不同类型的三极管的 β 值也不同,由制造商提供的规格给出。典型的 β 值从 100 到 1000 以内。 β 电流关系提供了一个快速判断需要多小的基极电流才能控制一个大得多的集电极电流的方法。越大的 β 值表示越大的信号放大能力。因为基极-发射极结本质上是一个二极管,必须在它两端加上足够的正偏电压才能使三极管导通(典型值 $V_{BE} = 0.7\text{ V}$)。PNP 三极管功能上与 NPN 类似,只不过电流和电压的方向刚好相反。

当三极管电路设计好以后,必须注意不能使基极-发射极结过载。和其他的二极管一样,当它处于正偏时表现出来的阻抗非常小。如果没有某种限流措施,三极管将会因为过热而烧毁。三极管常采用在它两个引脚或者三个引脚之间连接电阻来实现适当的电压偏置。图 2.3.5 显示的是一个工作在 DC 电流下的普通 NPN 三极管,其基极接有限流电阻 R_B ,集电极接有分压电阻 R_C 。发射极接地,使得基极电压 $V_B = 0.7\text{ V}$ 。 R_B 向基极提供电流用以控制集电极的电压 V_C 。当 R_B 的值变大时, V_C 随之变大,因为流出集电极的电流减小了,使得 R_C 上的压降也减小了。在这个例子中,基极电流 I_B 等于 $(5\text{ V} - V_B) / R_B = 0.43\text{ mA}$ 。假设 β 等于 100,那么集电极电流 I_C 就等于 43 mA ,并且 $V_C = 5\text{ V} - I_C R_C = 2.85\text{ V}$ 。

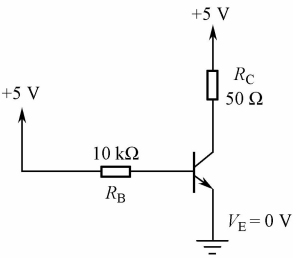


图 2.3.5 NPN 直流拓扑

三极管能驱动的电流大小受它本身的物理特性以及偏置方式的限制。物理上,三极管都有一个特定的最大功耗限制,超过了这个限制它就会因为过热而逐渐烧毁。在这个电路中,三极管的功耗约为 $V_{CE}I_C + V_{BE}I_B$ 。主要是由占总电流的绝大部分的、在集电极和发射极之间流过的电流决定。经过这种简化之后,得到这个三极管的功耗近似为 $2.85\text{ V} \times 43\text{ mA} = 123\text{ mW}$ 。

假设三极管在它的极限范围内正常工作,那么它的偏置方式决定了能够流入集电极的电流值上限。BJT 有三种工作状态:截止状态、放大状态和饱和状态。在截止状态下,三极管不导通,因为此时基极-发射极结要么处于反偏状态,要么处于正向偏压不够的状态。集电极上的电压此时处于最大值,并且基极集电极结处于反偏状态,因为没有电流流过, R_C 上也没有压降。当基极-发射极结正向偏置时,三极管导通并且 V_C 开始下降,此时三极管处于放大状态。只要基极-集电极结还处于反偏状态,上升的基极电流就会相应地引起集电极电流的上升,此时的三极管仍处于放大状态。如果 I_B 增大到某一个值,使得基极-集电极结变成正偏(上升的 I_C 导致 V_C 逐渐趋近于 V_E),三极管就进入了饱和状态,并且流过集电极的电流也不再变大。饱和状态并不会使三极管烧毁,但它却使 I_B 和 I_C 之间变成了非线性关系, β 也不再具有放大效果。如果 R_C 增大或者减小,则发生饱和现象所要求的 I_C 也就相应地增大或者减小。放大器电路必须避免饱和现象的出现,因为它会导致非线性。相反地,当在纯数字电路中工作时,只要没有超过功耗极限,三极管就不断在截止与饱和这两个状态之间切换。

2.3.5 Reading Materials

Diodes in Digital Applications

Not only can diode logic functions be implemented for power supply sharing or backup, they are equally applicable to implementing certain simple logic tasks on a circuit board. Diodes can implement both simple OR and AND functions and are useful when either a standard logic gate is unavailable or when the amplitude of the incoming signals violates the minimum or maximum input voltages of other components. Figure 2.3.6 shows diodes implementing two-input OR and AND functions. Pull-down (下拉) and pull-up (上拉) resistors are necessary for the OR and AND functions, respectively, because the diodes conduct only when forward biased. When both diodes are reverse biased, the circuit must be pulled to a valid logic state. The value of the resistors depends on the input current of the circuit being driven but ranges from 1 to 10 k Ω are common.

The pull-down resistor in the OR circuit maintains a default logic level of 0 when both inputs are also at logic 0. Both inputs must remain below $V_F = 0.7\text{ V}$ for the circuit to generate a valid logic 0 V level. When the input signals transition to logic 1, they must stabilize at a higher voltage that is sufficient to meet the minimum logic-1

input voltage of the driven circuit. The value of the pull-down resistor should be high enough to limit the power consumption of the circuit but low enough to create a voltage that is comfortably below the driven circuit's logic-0 threshold. A CMOS input has a much lower input current specification than a TTL input. A typical TTL input has a low-level input current of under 0.5 mA, and it should be kept well below 0.8 V for adequate margin. A 1 k Ω pull-down resistor would create less than a 0.5 V drop under these conditions. This may be adequate for some designs, or a more conservative approach could be taken by using a smaller resistance, perhaps 470 Ω . When either input rises to its logic-1 voltage, this will be reflected in the circuit's output minus a diode drop. This places a restriction on the input voltages; they cannot exceed the maximum input voltage of the driven circuit by more than a diode drop. However, the input voltages can violate the minimum input voltage specification, because the diodes will be reverse biased under these conditions and thereby prevent the circuit's output voltage from dropping below 0 V, or ground.

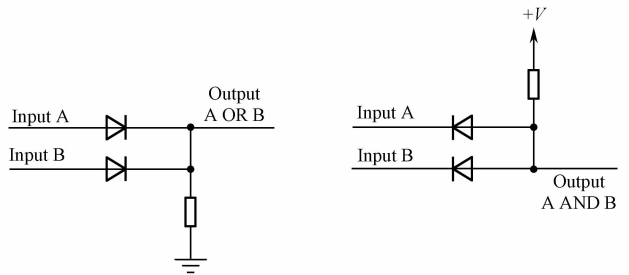


Figure 2. 3. 6 Diodes AND/OR.

2. 4 The Ideal of Op-amp(I)

2. 4. 1 Text

The design of amplifiers is normally most relevant in analog circuits such as those found in audio and RF communications. An amplifier is an analog circuit that outputs a signal with greater amplitude than what is presented to it at the input^[1]. Amplification is sometimes necessary in digital systems. Amplifiers are often found at interfaces where the weak signal from a transducer (e. g. , microphone or antenna) must be strengthened for sampling by an analog-to-digital converter. Even if a signal has sufficient amplitude, it may be desirable to scale it for better sampling resolution. For example, if an analog-to-digital converter accepts an input of 0 to 5 V and the incoming signal swings only between 0 and 3 V, 40 percent of the converter's resolution will be wasted. An amplifier can be used to scale the signal up to the full 5 V input range of the converter.

Solid-state amplifiers are constructed using transistors integrated onto a silicon chip or discrete transistors wired together on a circuit board. Amplifiers range greatly in complexity; complete AC analysis theory and its application to discrete amplifier design

are outside the scope of our discuss. However, the design of many general-purpose amplifiers is made easier by the availability of prebuilt components called operational amplifiers (op-amps). Op-amps are so common that they are considered to be basic building blocks in analog circuit design. An op-amp may contain dozens of transistors,

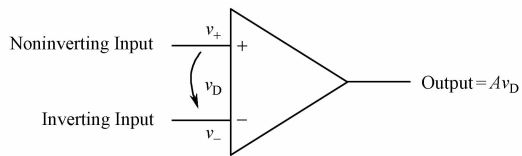


Figure 2.4.1 Op-amp graphical representation.

but its external interface consists of two differential inputs and an amplified output as shown in Figure 2.4.1. The positive, or noninverting, and the negative, or inverting, inputs form a differential voltage, $v_D = v_+ - v_-$, that

is amplified by a certain gain, A , at the output so that $v_O = Av_D = A(v_+ - v_-)$. When discussing the AC signals on which amplifiers operate, lower-case letters are used to indicate voltages and currents by convention to distinguish them from DC voltages and currents that have already been shown to use capital letters.

Many of the implementation details of how an AC signal is amplified within the op-amp are hidden from the circuit designer, requiring only an understanding of how the op-amp behaves from an external perspective^[2]. It is best to first explore an ideal opamp’s operation and then take into account the real-world deviations from the ideal model as necessary when designing a real circuit^[3]. An ideal op-amp has the following characteristics:

- Infinite input impedance. No current flows into or out of the inputs.
- Infinite open-loop gain. This may sound confusing, but most op-amp circuits employ feedback that reduces the infinite gain to the desired level. $A = \infty$ simplifies op-amp circuit analysis, as will soon be shown.
- Infinite bandwidth. The op-amp’s gain is constant across frequency from zero to infinity.
- Zero output impedance. The op-amp’s output will always be equal to Av_D regardless of the load being driven.

These fundamental assumptions provide the engineer with a very flexible amplifier component that can be customized by surrounding circuitry to suit a wide range of applications^[4]. Perhaps the first question that comes to mind is how an amplifier with infinite gain can be made useful^[5]. The trick is in creating a closed-loop circuit that provides feedback from output to input to control the gain of the overall circuit. Without a feedback path, an open-loop op-amp circuit would, in fact, exhibit very high gain to the point of grossly distorting most types of signals^[6]. Consider the basic noninverting closed-loop op-amp circuit in Figure 2.4.2. While the signal is injected into the positive input, the op-amp’s output feeds back to the negative input through the resistor

network formed by R_1 and R_2 .

$$\nu_- = \nu_0 \frac{R_1}{R_1 + R_2}$$

Knowing that $\nu_0 = A\nu_D = A(\nu_+ - \nu_-)$, this expression can be used to reveal a relationship between the input voltage, ν_1 , and ν_0 .

$$\nu_0 = A(\nu_+ - \nu_-) = A\nu_1 - A\left[\nu_0 \frac{R_1}{R_1 + R_2}\right]$$

This relationship can be simplified based on the assumption of infinite gain. Dividing both sides of the equation by $A = \infty$ causes the lone ν_0 term on the left-hand side to disappear because $\nu_0 \div \infty = 0$.

$$0 = \nu_1 - \frac{\nu_0 R_1}{R_1 + R_2}$$

Finally, the input and output terms of the equation can be separated onto separate sides of the equality to yield a final simplified relationship between ν_1 and ν_0 as follows:

$$\begin{aligned} \nu_1 &= \nu_0 \frac{R_1}{R_1 + R_2} \\ \nu_0 &= \nu_1 \left(\frac{R_1 + R_2}{R_1} \right) = \nu_1 \left[1 + \frac{R_2}{R_1} \right] \end{aligned}$$

This shows that, despite the ideal op-amp's infinite gain, the circuit's overall gain is easily quantifiable and controllable based on the two resistor values. A noninverting op-amp circuit can be used to scale up an incoming signal for a purpose such as that already mentioned; using all of the available resolution of an analog-to-digital converter. In this example, a transducer of some kind (e.g., temperature sensor or audio input device) creates a signal that ranges from 0 to 3 V, and the analog-to-digital converter that is sensing it has a fixed sampling range from 0 to 5 V. To take full advantage of the sampling range, it is desirable to apply a gain of 1.667 to the input signal so that it swings from 0 to 5 V. This is accomplished using the noninverting circuit shown in Figure 2.4.3. R_1 and R_2 are chosen arbitrarily as long as they satisfy the ratio $R_2 : R_1 = 2 : 3$ ^[7]. Values of 2.2 k Ω and 3.3 k Ω provide feedback in the desired ratio with relatively low maximum current draw on the order of 1 mA.

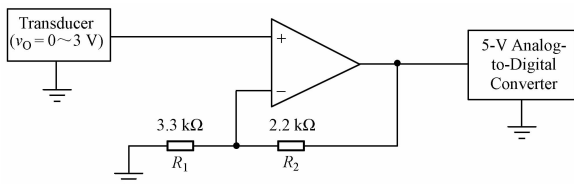


Figure 2.4.3 Scaling up an analog-to-digital converter input signal.

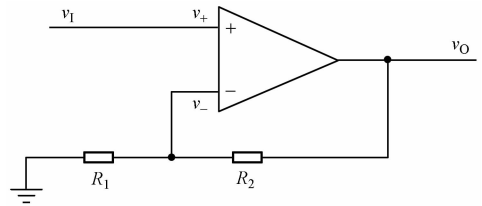


Figure 2.4.2 Noninverting op-amp circuit.

2. 4. 2 Specialized English Words

op-amp 运算放大器 (operation-amplifier 之缩写)	a differential voltage 差分电压
analog 模拟, 类比	AC signals 交流信号
circuits 电路	DC voltages 直流电压
amplifiers 放大器	impedance 阻抗, 全电阻
amplitude 振幅	open-loop 开环
relevant 相关的	gain 增益(系数)
RF 射频, 无线电频率(Radio Frequency 之缩写)	implementation 实现
weak signal 小信号	fundamental assumptions 基本假设
transducer 传感器, 变换器	feeds back 反馈
microphone 扩音器, 麦克风	quantifiable 可量化的
antenna 天线	resolution 分辨率
discrete 离散的	temperature sensor 温度传感器
transistors 晶体管	fixed sampling range 固定的采样范围
general-purpose 通用的	arbitrarily 任意的
interface 接口	approach 近似

2. 4. 3 Notes

[1]An amplifier is an analog circuit that outputs a signal with greater amplitude than what is presented to it at the input. 本句的主句“An amplifier is an analog circuit”为系表结构,其后为“that”引导的定语从句,修饰“analog circuit”,从句中有一“what”引起的名词性从句指代输入端的信号幅值。全句可译为“放大器是让输出信号的幅值比输入信号的幅值增大的模拟电路。”

[2]Many of the implementation details of how an AC signal is amplified within the op-amp are hidden from the circuit designer, requiring only an understanding of how the op-amp behaves from an external perspective. 本句主句的主语为“Many of the implementation details”,谓语为“are hidden from”。句中“of how ... op-amp”为介词短语结构,用做主语的后置定语,而介词“of”后面的“how an AC signal is amplified within the op-amp”则是名词从句,用做“of”的介词宾语。句末“requiring”以下为现在分词短语,作为主句的状语,表示结果。全句可译为“当直接采用运算放大器设计电路时,我们可以忽略那些交流信号在运算放大器内是如何被放大的细节,只需要了解运算放大器的外部特性就可以了。”

[3]It is best to first explore an ideal op-amp’s operation and then take into account the real-world deviations from the ideal model as necessary when designing a real

circuit. 这是一个并列复合句,“explore”和“take”为两个并列谓语动词,句末“when designing a real circuit”为条件状语,可以认为是“when we are designing”,省略了“we are”。全句可译为“采用运算放大器设计时,最好一开始先采用理想模型,再考虑实际模型与理想模型的偏差。”

[4] These fundamental assumptions provide the engineer with a very flexible amplifier component that can be customized by surrounding circuitry to suit a wide range of applications. 这是一个结构比较清晰的主从复合句。“that can be customized…”为定语从句,修饰“amplifier component”。全句可译为“上述基本假设可以提供给设计工程师一个非常灵活的器件,他们可以根据周围电路来设计运算放大器以适应不同的需要。”

[5] Perhaps the first question that comes to mind is how an amplifier with infinite gain can be made useful. 本句不长,但结构上有一定的复杂性,主句为系表动词+表语句型。主语为“the first question”,联系动词为“is”,“how …”为名词性表语从句。“that comes to mind”为定语从句,修饰主句“the first question”。全句可译为“也许映入我们脑海的第一个问题就是如何让一个增益无穷大的运算放大器发挥作用。”

[6] Without a feedback path, an open-loop op-amp circuit would, in fact, exhibit very high gain to the point of grossly distorting most types of signals. 本句为一简单句,但附属成分较多。“Without a feedback path”和“in fact”均为介词短语做状语,“to the point of …”亦为介词短语做状语,表示程度。“distorting”为动名词做介词“of”的宾语。全句可译为“事实上,如果没有反馈回路,运算放大器的开环增益将会非常大,造成绝大多数的输入信号严重扭曲失真。”

[7] R_1 and R_2 are chosen arbitrarily as long as they satisfy the ratio $R_2 : R_1 = 2 : 3$. 本句为一主从复合句,主句为“ R_1 and R_2 are chosen arbitrarily”,“as … as …”构成比较句型,前一“as”为副词,修饰“long”,后一“as”为连接词,引起从句“they satisfy …”。全句可译为“ R_1 和 R_2 的阻值可以任意确定,只要它们满足 $R_2 : R_1 = 2 : 3$ 的比例即可。”

2.4.4 Reference Translation

理想运算放大器(I)

放大器的设计在模拟电路设计中是非常普遍的,这些电路在音频和射频通信中经常使用。放大器是让输出信号的幅值比输入信号的幅值增大的模拟电路。在数字电路中有时也要用到信号放大。例如,接口中的放大器将传感器(如麦克风或天线)产生的微弱信号增强后,再传输到模/数转换的采样电路。即使一个信号有足够的幅值,我们也希望将它放大到一定比例以获得更好的采样分辨率。例如,一个模数转换器接收 $0\text{ V} \sim 5\text{ V}$ 的输入信号,而输入信号的范围是 $0\text{ V} \sim 3\text{ V}$,这样 40% 的转换器分辨率浪费了。因此,我们可以将 $0\text{ V} \sim 3\text{ V}$ 的输入信号放大到转换器的 5 V 的满程范围。

固态放大器既可以采用单片集成晶体管的方式实现,也可以采用在电路板上焊接分离元件的方式实现。放大器涉及的电路范围很广。完整的交流分析理论和将其用于分离

元件实现放大器的设计超出了我们的讨论范围。但是很多通用放大器的设计,如果采用称为运算放大器的预制器件来进行,就变得容易了。运算放大器的应用是如此普遍,以至于它可以被视为模拟电路的基本构成模块。

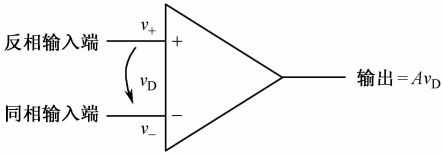


图 2.4.1 基本反相放大电路

一个运算放大器可能包含许多晶体管,但它的外部接口只包含两个不同的输入信号和一个放大信号输出,如图 2.4.1 所示。正输入(或同相输入 v_+)和负输入(或反相输入 v_-)存在一个电压差 $v_D = v_+ - v_-$ 。假设放大器增益为 A , 输出则为 $v_o = A v_D = A(v_+ - v_-)$, 当讨论交流信号时,我们按规定用小写字母表示电流电压,以区别大写字母表示的直流电压和电流。

当直接采用运算放大器设计电路时,我们可以忽略那些交流信号在运算放大器内是如何被放大的细节,只需要了解运算放大器的外部特性就可以了。采用运算放大器设计时,最好一开始先采用理想模型,再考虑实际模型与理想模型的偏差。运算放大器包含的特性主要如下:

- 输入电阻无穷大:无电流进出输入端口。
- 开环增益无穷大:这听起来有些混淆,但大多数运算放大器电路采用了反馈回路,将增益从无穷大减少到预期水平。下面会看到, $A = \infty$ 可以简化运算放大器电路的分析。
- 带宽无穷大:输入频率从 0 到无穷大,运算放大器的增益均为常数。
- 输出阻抗零:不管驱动负载多大,运算放大器的输出将永远为 $A v_D$ 。

上述基本假设可以提供给设计工程师一个非常灵活的器件,他们可以根据周围电路来设计运算放大器以适应不同的需要。也许映入我们脑海的第一个问题就是如何让一个增益无穷大的运算放大器发挥作用。诀窍就是建立闭环电路,通过采用从输出到输入的反馈回路来控制全电路的增益。事实上,如果没有反馈回路,运算放大器的开环增益将会非常大,造成绝大多数的输入信号严重扭曲失真。图 2.4.2 显示的是一个基本的反相闭环放大电路。输入信号连接到正相端,运算放大器的输出通过 R_1 和 R_2 构成的电阻网络反馈到反相输入端。

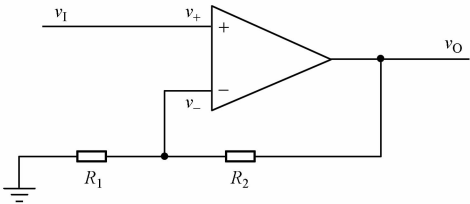


图 2.4.2 同相放大电路

$$v_- = v_o \frac{R_1}{R_1 + R_2}$$

根据 $v_o = A v_D = A(v_+ - v_-)$, 可将上面的表达式进行变换,进一步揭示出输入电压 v_i 和输出电压 v_o 之间的关系:

$$v_o = A(v_+ - v_-) = A v_i - A \left[v_o \frac{R_1}{R_1 + R_2} \right]$$

基于增益无穷大的假设,输入电压 v_i 和输出电压 v_o 之间的关系可以进一步简化。上

述公式的两边同时除以 $A = \infty$,公式左边的单一项 ν_o 除以无穷大结果为 0,则公式进一步变化如下:

$$0 = \nu_i - \frac{\nu_o R_1}{R_1 + R_2}$$

最后,将输出信号 ν_o 和输入信号 ν_i 分别放到公式左右两边,我们可以得到公式最终的简化形式如下:

$$\begin{aligned} \nu_i &= \nu_o \frac{R_1}{R_1 + R_2} \\ \nu_o &= \nu_i \left(\frac{R_1 + R_2}{R_1} \right) = \nu_i \left[1 + \frac{R_2}{R_1} \right] \end{aligned}$$

通过这个例子可以看出,尽管理想运算放大器的增益无穷大,但是只要通过两个电阻值我们可以很容易量化和改变电路的增益。一个非反相运算放大器可以用来将输入信号按比例地增大,从而实现前述的目的;充分利用模数转换器的分辨率。在这个例子中,某种传感器(如温度传感器或音频输入设备等)生成 $0\text{ V} \sim 3\text{ V}$ 的信号,而用这样的模数转换器再进行传感则可具有 $0\text{ V} \sim 5\text{ V}$ 的固定采样范围。为了充分利用 $0\text{ V} \sim 5\text{ V}$ 的采样范围,我们希望通过获得对输入信号的 1.667 倍增益,从而实现 $0\text{ V} \sim 5\text{ V}$ 的输出范围。这可以用图 2.4.3 所示的非反相电路实现。其中 R_1 和 R_2 的阻值可以任意确定,只要它们满足 $R_2 : R_1 = 2 : 3$ 的比例即可。采用 $2.2\text{ k}\Omega$ 和 $3.3\text{ k}\Omega$ 的电阻就可以获得预期的反馈比例,而电阻中最小电流处在相当小的 1 mA 数量级上。

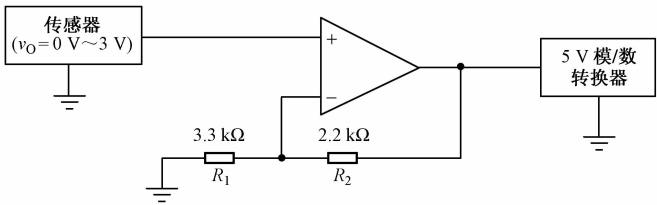


图 2.4.3 比例放大模数转换电路

2.4.5 Reading Materials

Active Filters(有源滤波器)

Active filters perform the same basic frequency passing and blocking function as passive filters, but they can simultaneously amplify the signal to form a filter that has unity or higher gain. This is in contrast to passive filters that achieve less than unity gain because of finite losses inherent in the components from which they are constructed. Op-amps can be used to implement active filters as long as their gain-bandwidth product is not exceeded. Figure 2.4.4 shows familiar first-order low pass and highpass active filters implemented with op-amps. These simple filters buffer a passive

filter with a noninverting op-amp stage. In this example, the configurations are for unity gain, although higher gains are possible. Because of the high input resistance of an op-amp, there is little signal loss through the series elements while operating in the passband(通频带). Unlike a passive filter whose characteristics are influenced by the load being driven, the op-amp isolates the load from the filter elements.

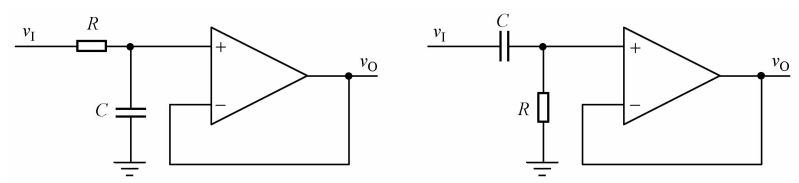


Figure 2. 4. 4 First-order active filters.

Filter design is a topic in electrical engineering that can get quite complex when very specific and demanding frequency response characteristics are necessary. Active filters add to this complexity as a result of nonideal op-amp characteristics. Although a complete discussion of filter design is outside the scope of this text, certain filtering tasks can be accomplished by drawing on a basic familiarity with common circuits. A common second-order topology used to implement active filters is the Sallen-Key filter (增益为 1 的巴特沃斯滤波器). Sallen-Key lowpass(低通) and highpass(高通) filters are implemented with two resistors and two capacitors each for unity gain in the passband, as shown in Figure 2. 4. 5. If higher gains are desired, two resistors can be added per the standard noninverting amplifier circuit topology. As with a passive second-order filter, the frequency response curve falls off at 40 dB per decade beyond the cut-off frequency.

$$f_c = \frac{1}{2\pi \sqrt{R_1 R_2 C_2}}$$

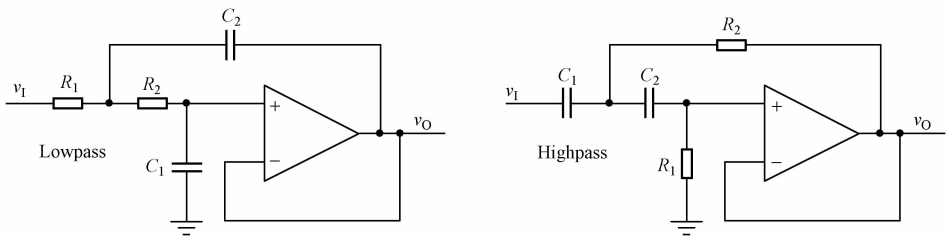


Figure 2. 4. 5 Sallen-Key second-order active filters.

2.5 The Ideal of Op-amp(II)

2.5.1 Text

If a circuit such as this is implemented in reality, it would be helpful to understand the effect of finite op-amp gain on the circuit's overall gain to ensure that the circuit operates as intended. To do so, we can turn back to the previous derivation of op-amp gain. Instead of dividing the equation by $A = \infty$ early in the derivation, the input and output terms are immediately separated and re-expressed as follows:

$$\begin{aligned} \nu_o + A \left[\frac{\nu_o R_1}{R_1 + R_2} \right] &= A \nu_1 = \nu_o \left[1 + A \frac{R_1}{R_1 + R_2} \right] \\ \nu_o &= \nu_1 \frac{A}{1 + A \frac{R_1}{R_1 + R_2}} = \nu_1 \frac{A \frac{R_1 + R_2}{R_1}}{\frac{R_1 + R_2}{R_1} + A} = \nu_1 \frac{A \left[1 + \frac{R_2}{R_1} \right]}{1 + \frac{R_2}{R_1} + A} \end{aligned}$$

This relationship is more complex than the ideal case, but it can be seen that, as A approaches infinity, the expression simplifies to that which has already been presented. A wide variety of op-amps are manufactured with differing gains and electrical characteristics. One of the most common opamps is the LM741, a device that has been around for decades. The LM741 has a minimum voltage gain of 20 V/mV, or $A = 20,000$ V/V. For small circuit gains where R_2/R_1 is much less than A , the LM741 will provide an overall circuit gain that is extremely close to the ideal. Using the previous example, the gain expression becomes

$$\nu_o = \nu_1 \frac{20000 \left[1 + \frac{2.2}{3.3} \right]}{1 + \frac{2.2}{3.3} + 20000} = \nu_1 \frac{33333}{20001.67} = \nu_1 1.6665$$

It can be observed that, for a real-world op-amp gain of much less than infinity, the ideal gain expression for an op-amp provides a very accurate calculation^[1]. As the gain desired from the circuit is increased, the denominator of the nonideal gain expression will increase as well, causing greater divergence between ideal and nonideal calculations for a given op-amp gain specification. Of course, the LM741 is not the only op-amp available. Newer and more advanced designs are readily available with gains an order of magnitude higher than that of the LM741.

The minimum gain achievable by the noninverting op-amp circuit is 1, or unity gain, when $R_2 = 0 \Omega$. There are instances in which a unity-gain buffer stage is desired. An example is the need to isolate a weak driver from a heavy load^[2]. While an ideal op-amp has infinite input impedance, a real op-amp has very high input impedance.

Consequently, even a nonideal op-amp will present a light load to a driver. And while a real op-amp has nonzero output impedance, it will be much lower than the weak driver being isolated. As shown in Figure 2.5.1, a unity-gain buffer is constructed by directly feeding the output back to the negative input. It can be observed from the previous circuit that when $R_2 = 0 \, \Omega$, R_1 becomes superfluous.

A limitation of the noninverting op-amp circuit is that the minimum gain achievable is 1. When a gain of less than 1 is desired, a slightly different circuit topology is used: the inverting configuration. As shown in Figure 2.5.2, the noninverting input is grounded, and the signal is injected into the negative input through R_1 . As before, R_2 forms the feedback loop that stabilizes the circuit's overall gain.

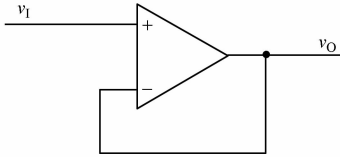


Figure 2.5.1 Unity-gain buffer.

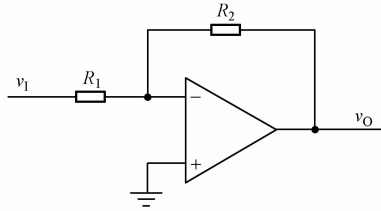


Figure 2.5.2 Op-amp inverting close-loop circuit.

The inverting configuration's relationship between v_I and v_O can be derived using the resistor divider method shown previously for the noninverting circuit. Op-amp circuits can also be analyzed with an alternative method that provides a slightly different view of their operation. In some cases, mathematical analysis is made somewhat easier using one of the two methods. The alternative method uses the assumption of infinite gain to declare that the differential input voltage, v_D , equals zero. If $A = \infty$ and $v_O = A v_D$, it follows that $v_D = 0$ for a finite v_O . This assumption leads to an implied virtual short circuit between the op-amp's two input terminals. If $v_D = 0$, $v_+ = v_-$. The virtual short circuit tells us that if a voltage is applied at the positive terminal, it will appear at the negative terminal as well, and vice versa. Therefore, rather than expressing v_- as a resistor divider between v_O and v_I , each portion of the circuit can be analyzed separately. In such a simple circuit, this concept may not seem to have much advantage. However, the analysis of more complex op-amp circuits can benefit from this approach.

To demonstrate circuit analysis using the virtual short circuit approach, we begin by knowing that $v_- = 0 \text{ V}$, because the positive terminal is grounded. Therefore, the voltage drop across R_1 is known by inspection, and its current, i_1 , is simply v_I/R_1 . We know from basic circuit theory that current cannot just disappear. Assuming that the op-amp has infinite input impedance, all of i_1 must flow toward v_O . Hence, $i_2 = -i_1$. The output voltage can now be determined using Ohm's law to show that the overall

gain is controlled by the resistors when an ideal op-amp is used.

$$v_O = v_- + i_2 R_2 = 0 - i_1 R_2 = -v_I \frac{R_2}{R_1}$$

The inverting circuit can be designed with arbitrary gains of less than 1. However, both a positive and negative voltage supply are required to enable the op-amp to drive both positive and negative voltages^[3]. If a signal with a voltage range from 0 V to 3 V is applied to an inverting circuit with a gain of 0.8, the op-amp will generate an output signal from -2.4 V to 0 V. In some situations, this may be undesirable because of the requirement imposed by processing negative voltages. Fortunately, op-amps are very flexible, and the inverting configuration can be biased to center the output signal about a nonzero DC level. Consider the circuit in Figure 2.5.3. Rather than grounding the positive input, a bias voltage is applied.

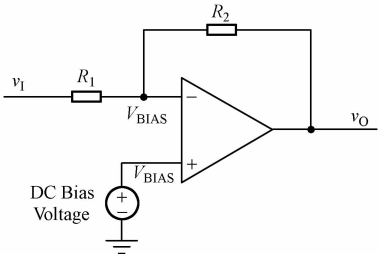


Figure 2.5.3 Biased inverting op-amp circuit.

2.5.2 Specialized English Words

isolated	隔离	noninverting input	同相输入
weak	微弱的	buffer	缓冲
heavy load	重负载	feedback loop	反馈环
impedance	阻抗		

2.5.3 Notes

- [1]It can be observed that, for a real-world op-amp gain of much less than infinity, the ideal gain expression for an op-amp provides a very accurate calculation. 这是一主从复合句,“It”为形式主语,代替真正的主语,即名词从句“the ideal gain expression for an op-amp provides a very accurate calculation”,“for a ... infinity”为介词短语,修饰“provides”,表示原因或条件。全句可译为“通过上述实例我们可以看出,尽管实际运算放大器的增益远小于无穷大,但是采用理想模型仍能够提供非常精确的计算结果。”
- [2]An example is the need to isolate a weak driver from a heavy load. 这是个简单句,“to isolate ... from ...”为不定式短语做定语,修饰“the need”。全句可译为“一个例子就是需要将弱信号与重负载隔离开来(的情况)。”
- [3]However, both a positive and negative voltage supply are required to enable the op-amp to drive both positive and negative voltages. 此句为简单句,注意动词“are”为复数形式,因“supply”为不可数名词,此处为复数(a positive and negative)。全句可译为“但是为了使放大器能够输出正电压和负电压,必需提供正负两种电源。”

2. 5. 4 Reference Translation

理想运算放大器(II)

如果上述的电路得以实现，将有助于我们理解如何利用有限增益的运算放大器实现预期的电路增益。为了做到这一点，让我们回到前面对运算放大器增益的推导。不将等式两边像前面那样除以 $A = \infty$ ，而是直接就对输入项与输出项进行分离如下：

$$\begin{aligned} \nu_o + A\left[\frac{\nu_o R_1}{R_1 + R_2}\right] &= A\nu_i = \nu_o\left[1 + A\frac{R_1}{R_1 + R_2}\right] \\ \nu_o &= \nu_i \frac{A}{1 + A\frac{R_1}{R_1 + R_2}} = \nu_i \frac{A\frac{R_1 + R_2}{R_1}}{\frac{R_1 + R_2}{R_1} + A} = \nu_i \frac{A\left[1 + \frac{R_2}{R_1}\right]}{1 + \frac{R_2}{R_1} + A} \end{aligned}$$

通过上述公式，我们可以看出实际运算放大器的输出电压与输入电压之间的关系比理想模型的要复杂得多。但是当 A 趋于无穷大时，上述公式就可以简化成前面所得到的公式。已经生产出许各种各样不同增益和电气特性的运算放大器，其中 LM741 是最常用的运算放大器之一，已经活跃了几十年。LM741 具有 20 V/mV 的最小电压增益，即 $A=20000$ V/V。为了得到较小的电路增益，要求 $R_2 \div R_1$ 的值远小于 A ，这时 LM741 能够提供非常接近于理想值的电路增益。应用前述的实例，增益表达式变为

$$\nu_o = \nu_i \frac{20\,000\left[1 + \frac{2.2}{3.3}\right]}{1 + \frac{2.2}{3.3} + 20\,000} = \nu_i \frac{33\,333}{20\,001.67} = \nu_i 1.6665$$

通过上述实例我们可以看出，尽管实际运算放大器的增益远小于无穷大，但是采用理想模型仍能够提供非常精确的计算结果。然而，当对电路的预期增益要求增大时，导致输入电压 ν_i 和输出电压 ν_o 之间的关系公式的分母也随之增大，从而导致理想模型与非理想模型之间偏差增大。当然 LM741 不是唯一可用的运算放大器。在实际电路设计中我们可以采用比 LM741 更新、更先进的、放大增益量级比 LM741 的大一个数量级的运算放大器。

对于同相运算放大器电路而言，可获得的最小增益是 1，这也被称为单位增益，此时 $R_2 = 0\,\Omega$ 。有些情况下，需要采用单位增益缓冲级。一个例子就是需要将弱信号与重负载隔离开来(的情况)。理想运算放大器的输入阻抗为无穷大，实际的运算放大器的输入阻抗也是非常大的。这样，即使是实际运算放大器对于驱动电路而言也表现为一个很轻的负载；而理想运算放大器的输出阻抗为 0，将远比所隔离的驱动电路的输出阻抗低得多。如图 2. 5. 1 所示，单位增益运算放大器可以通过将输出直接反馈到反向输入端构成。从前面的电路可以看出，当 $R_2 = 0\,\Omega$ 时， R_1 可以略去。

同相运算放大器的一个局限是最小增益为 1。要想获得小于 1 的增益，可采用反相结构稍有不同电路，如图 2. 5. 2 所示，同相输入端接地，输入信号通过电阻 R_1 连接到反相输入端。同以往一样， R_2 的作用是形成反馈回路，从而确定了电路增益。

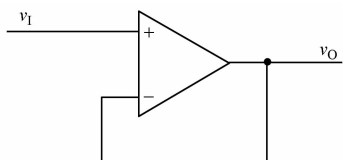


图 2.5.1 单位增益缓冲

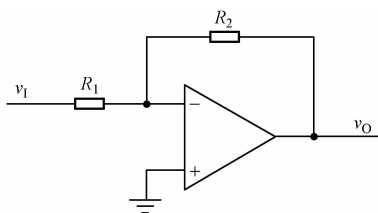


图 2.5.2 闭环反相放大电路

反相电路的输入 v_1 与输出从 v_O 之间的关系同样可以采取与同相电路类似的电阻相除的方法进行分析。我们也可以采取另外一种方法从另外一个稍有不同的角度来分析运算放大器的行为。在某些情况下,采用这两种数学方法进行分析多少要简单些。这种方法基于电路增益无穷大的假设因而差分输入电压差 v_D 为 0。如果 $A = \infty$, $v_O = A v_D$, 对于有限的 v_O , 可以得到 $v_D = 0$ 。从这种假设可以推出放大器的两个输入端口虚短。如果 $v_D = 0$, 则 $v_+ = v_-$ 。由虚短理论可知,加在正相端脚 v_+ 的电压,也同样加在了反相端 v_- , 反之亦然。当然,除了可以将 v_O 与 v_1 之间的比例关系利用电阻相除来表达外,电路的每一个端口均可以进行独立的分析。在这个简单的例子中,这个概念起的作用似乎并不大,但是分析一些更加复杂的运算放大器时,运用这个概念将会大有裨益。

为了用虚短理论来进行电路分析,我们首先假设 $v_- = 0$ V, 此时正相端口接地。经观察,可以知道 R_1 的电压降,而通过它的电流 i_1 , 可以很容易地由 v_1/R_1 得到。根据基本电流理论,我们知道电流不会消失,又假设运算放大器的输入阻抗为无穷大,则所有的 i_1 电流均流向输出端 v_O , 因此 $i_2 = -i_1$ 。根据理想的放大器模型,我们可以用欧姆定律获得输出电压,即全电路的增益由电阻决定:

$$v_O = v_- + i_2 R_2 = 0 - i_1 R_2 = -v_1 \frac{R_2}{R_1}$$

我们可以设计增益小于 1 的任何反向电路,但是为了使放大器能够输出正电压和负电压,必需提供正负两种电源。设反向电路增益为 0.8, 输入信号的电压范围为 0 V 到 3 V, 则输出信号的电压范围为 -2.4 V 到 0 V。在一些应用场合中,我们不希望处理负电压。幸运的是,运算放大器非常灵活。我们可以通过在反相电路构造上增加一个偏置电压,从而使输出信号处在以一个非 0 的直流电平为中间值的范围里。分析一下图 2.5.3 所示的电路,其中正向输入不再接地,而是接一个偏置电压。

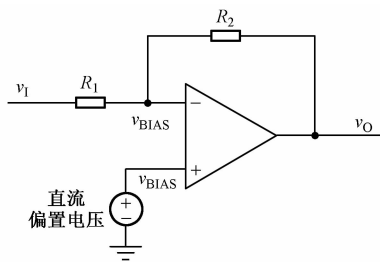


图 2.5.3 偏置放大电路

2.6 Boolean Algebra

2.6.1 Text

All digital systems are founded on logic design. Logic design transforms algorithms and processes conceived by people into computing machines. A grasp of digital logic is crucial to the understanding of other basic elements of digital systems, including microprocessors. This chapter addresses vital topics ranging from Boolean algebra to synchronous logic to timing analysis with the goal of providing a working set of knowledge that is the prerequisite for learning how to design and implement an unbounded range of digital systems.

Boolean algebra is the mathematical basis for logic design and establishes the means by which a task's defining rules are represented digitally^[1]. The topic is introduced in stages starting with basic logical operations and progressing through the design and manipulation of logic equations. Binary and hexadecimal numbering and arithmetic are discussed to explain how logic elements accomplish significant and practical tasks.

With an understanding of how basic logical relationships are established and implemented, the discussion moves on to explain flip-flops and synchronous logic design^[2]. Synchronous logic complements Boolean algebra, because it allows logic operations to store and manipulate data over time. Digital systems would be impossible without a deterministic means of advancing through an algorithm's sequential steps. Boolean algebra defines algorithmic steps, and the progression between steps is enabled by synchronous logic.

Machines of all types, including computers, are designed to perform specific tasks in exact well defined manners. Some machine components are purely physical in nature, because their composition and behavior are strictly regulated by chemical, thermodynamic, and physical properties. For example, an engine is designed to transform the energy released by the combustion of gasoline and oxygen into rotating a crankshaft. Other machine components are algorithmic in nature, because their designs primarily follow constraints necessary to implement a set of logical functions as defined by human beings rather than the laws of physics. A traffic light's behavior is predominantly defined by human beings rather than by natural physical laws^[3]. This book is concerned with the design of digital systems that are suited to the algorithmic requirements of their particular range of applications. Digital logic and arithmetic are critical building blocks in constructing such systems.

An algorithm is a procedure for solving a problem through a series of finite and specific steps. It can be represented as a set of mathematical formulas, lists of sequential

operations, or any combination thereof. Each of these finite steps can be represented by a Boolean logic equation. Boolean logic is a branch of mathematics that was discovered in the nineteenth century by an English mathematician named George Boole^[4]. The basic theory is that logical relationships can be modeled by algebraic equations. Rather than using arithmetic operations such as addition and subtraction, Boolean algebra employs logical operations including AND, OR, and NOT. Boolean variables have two enumerated values: true and false, represented numerically as 1 and 0, respectively.

The AND operation is mathematically defined as the product of two Boolean values, denoted A and B for reference. Truth tables are often used to illustrate logical relationships as shown for the AND operation in Table 2. 6. 1. A truth table provides a direct mapping between the possible inputs and outputs. A basic AND operation has two inputs with four possible combinations, because each input can be either 1 or 0 — true or false. Mathematical rules apply to Boolean algebra, resulting in a nonzero product only when both inputs are 1.

Table 2. 6. 1 AND Operation Truth Table.

A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

Summation is represented by the OR operation in Boolean algebra as shown in Table 2. 6. 2. Only one combination of inputs to the OR operation result in a zero sum: $0+0=0$.

Table 2. 6. 2 OR Operation Truth Table.

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

AND and OR are referred to as binary operators, because they require two operands. NOT is a unary operator, meaning that it requires only one operand. The NOT operator returns the complement of the input: 1 becomes 0, and 0 becomes 1. When a variable is passed through a NOT operator, it is said to be inverted.

When logical functions are converted into circuits, graphical representations of the seven basic operators are commonly used. In circuit terminology, the logical operators are called gates. Figure 2. 6. 1 shows how the basic logic gates are drawn on a circuit diagram. Naming the inputs of each gate A and B and the output Y is for reference only;

any name can be chosen for convenience. A small bubble is drawn at a gate’s output to indicate a logical inversion.

More complex Boolean functions are created by combining Boolean operators in the same way that arithmetic operators are combined in normal mathematics^[5]. Parentheses are useful to explicitly convey precedence information so that there is no ambiguity over how two variables should be treated. A Boolean function might be written as

$$Y = (AB + \overline{C} + D) \& \overline{E \oplus F}$$

This same equation could be represented graphically in a circuit diagram, also called a schematic diagram, as shown in Figure 2.6.2. This representation uses only two-input logic gates. As already mentioned, binary operators can be chained together to implement functions of more than two variables.

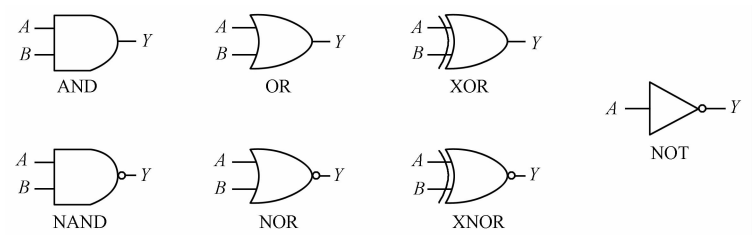


Figure 2.6.1 Graphical representation of basic logic gates.

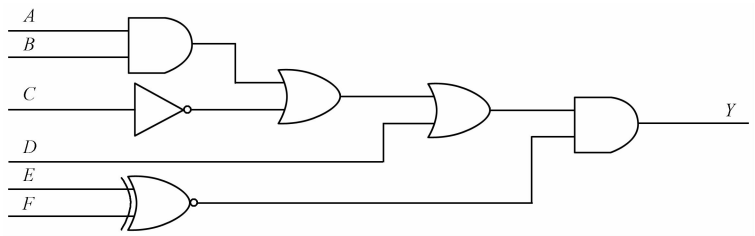


Figure 2.6.2 Schematic diagram of logic function.

An alternative graphical representation would use a three-input OR gate by collapsing the two-input OR gates into a single entity.

2.6.2 Specialized English Words

algorithms	算法	unary	一元, 单目
task	任务, 作业	logical functions	逻辑函数
accomplish	完成, 实现	graphical representations	图形表示法
established	建立	terminology	术语
over time	过去	parentheses	圆括号
components	元件	explicitly	明确的

2. 6. 3 Notes

[1] Boolean algebra is the mathematical basis for logic design and establishes the means by which a task's defining rules are represented digitally. 本复合句的主句有两个并列谓语动词, “is” 和 “establishes”。“by which a task's defining rules are represented digitally”是定语从句, 修饰 “means”。全句可译为“布尔代数是逻辑设计的数学基础, 并建立了用数字系统表示一项任务的规则的方法。”

[2] With an understanding of how basic logical relationships are established and implemented, the discussion moves on to explain flip-flops and synchronous logic design. 本复合句的主句为 “the discussion moves on to...”, “With an understanding of ...” 为介词短语做状语, 其中 “how basic logical relationships are established and implemented” 为名词从句, 做 “of” 的介词宾语。注意这两个名词从句有两个并列的被动形式的谓语动词 “are established and implemented”。全句可译为 “在了解了基本逻辑关系的建立和实现的基础上, 接着讨论了触发器和同步逻辑设计。”

[3] A traffic light's behavior is predominantly defined by human beings rather than by natural physical laws. 这是一个简单句, “rather than” 在此表示否定, 全句可译为 “交通灯的行为就不是由自然物理法则定义的, 而是由人预先定义的。”

[4] Boolean logic is a branch of mathematics that was discovered in the nineteenth century by an English mathematician named George Boole. 这是一个典型的复合句。“Boolean logic is a branch of mathematics” 为主句, 其表语 “a branch of mathematics” 带有 “that” 引导的从句做定语。句末 “named George Boole” 为过去分词短语, 做 “English mathematician” 的定语。全句可译为 “布尔代数是数学的一个分支, 在 19 世纪由英国数学家乔治·布尔率先提出。”

[5] More complex Boolean functions are created by combining Boolean operators in the same way that arithmetic operators are combined in normal mathematics. 这是一个典型的复合句, 结构较清楚。“that arithmetic operators are combined in normal mathematics” 为 “the same way” 的定语。全句可译为 “可以用与把算术操作符组合成数学表达式相同的方法来组建更复杂的布尔函数。”

2. 6. 4 Reference Translation

布 尔 代 数

所有的数字系统都是以逻辑设计为基础的。逻辑设计将人们提出的算法和处理转化到计算机中。为了更好地理解包含微处理器在内的数字系统的其他基本组件, 掌握数字逻辑设计是至关重要的。本章主要介绍一些数字系统设计和实现的基本知识, 如布尔代数、同步逻辑以及时序分析等, 这些有用的知识是学习如何设计和实现浩如烟海的数字系统的前提。

布尔代数是逻辑设计的数学基础,并建立了用数字系统表示一项任务的规则的方法。本章从逻辑公式的设计和处理开始来逐步介绍基本的逻辑操作及扩展。本章通过讨论数的二进制和十六进制表示和运算来解释逻辑单元是如何完成有实际意义的任务的。

在了解了基本逻辑关系的建立和实现的基础上,讨论了触发器和同步逻辑设计。同步逻辑是布尔代数的补充,它允许按时序对存储和处理数据进行逻辑操作。但是如果缺少一个明确的算法顺序,数字系统也无法起作用。布尔代数定义算法步骤,同步逻辑则控制算法进行。

包括计算机在内的所有机器都是以严格定义好的行为来执行一个特定的任务的。有些机器部件本质上是纯物理的,它们的组成和行为受化学特性、热力学特性以及物理特性的控制。例如,一个发动机被设计成将汽油和氧气混合燃烧释放的能量转换为曲轴的旋转。另外一些机器部件本质上是纯算法的,它们主要被设计来在一定约束条件下完成由人而不是由物理法则定义的一系列逻辑功能。例如,交通灯的行为就不是由自然物理法则定义的,而是由人预先定义的。本书关注的是那些能够满足特定应用的算法需求的数字系统设计。数字逻辑和运算正是这些数字系统的关键组成部分。

算法就是通过一系列有限的特定步骤解决问题的过程。它可以表示为一组数学表达式和时序操作,或者是它们的组合操作。其中每一确定步骤都可由一个布尔逻辑表达式表示。布尔代数是数学的一个分支,在 19 世纪由英国数学家乔治·布尔率先提出。它的基本理论是通过代数方程的方法来建立逻辑关系的模型。布尔代数采用的是与(AND)、或(OR)和非(NOT)等逻辑运算,而不是采用加、减等代数运算。布尔变量取值只有两个枚举值:真和假,用数字表示则分别为 1 和 0。

数学上的逻辑与(AND)运算被定义为两个用 A 和 B 表示的逻辑自变量的乘积。通常采用真值表来说明逻辑关系,表 2.6.1 给出的是与操作的真值表。真值表能够提供各种输入与输出之间直接的对应关系。一个基本与操作包含两个输入。由于每一个输入均有两种可能的取值:1 或 0,即真或假,则两个输入共有 4 种可能组合。如果把数学规则应用到与操作上,结果只有当两个输入都是 1 时才能得到非 0 的输出,其他情况输出均为 0。

表 2.6.1 AND 运算真值表

A	B	$A \text{ AND } B$
0	0	0
0	1	0
1	0	0
1	1	1

在布尔代数中“和”用 OR 运算来表示。OR 操作的真值表见表 2.6.2,其中只有一种输入组合的输出结果为 0,即 $0+0=0$ 。

表 2.6.2 OR 运算真值表

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

与(AND)运算和或(OR)运算都是二元算符,需要两个操作数。非(NOT)运算是一元算符,这意味着它只需要一个操作数。NOT 运算得到输入值的反码,即输入为 1,输出为 0 ,或输入为 0,输出为 1。因此当一个变量经过 NOT 运算,就被称为取反。

将逻辑函数转换成具体电路时,经常使用到七种基本逻辑的图形表示。在电路术语中这些逻辑运算符被称做门。一般情况下,电路图中的基本逻辑门的图形表示如图 2.6.1所示,其中输入称为为 A 和 B,输出为 Y 仅是一个举例,可根据设计者的方便改用任何名称。这里需要说明的是,门输出上加一个小圆圈表示为逻辑反相。

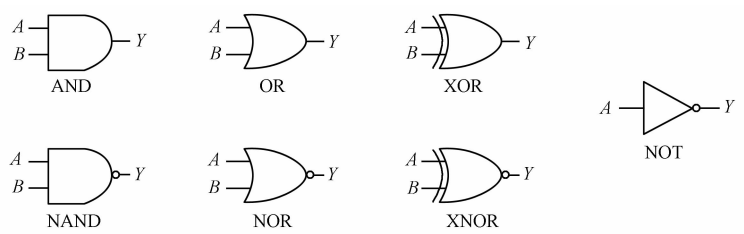


图 2.6.1 基本逻辑门的图形表示

可以用与把算术运算符组合成数学表达式相同的方法来组建更复杂的布尔函数。圆括号在准确表达优先信息的时是非常有用的,所以在处理两个变量时,不会造成含糊不清。一个布尔函数可能表示成

$$Y = (AB + \overline{C} + D) \& \overline{E \oplus F}$$

同一函数可以用电路图进行等效表达,称为符号图,如图 2.6.2 所示。这个电路图全都是采用二输入逻辑门。前已述及,二进制算符可以链接起来实现大于二输入的逻辑函数。

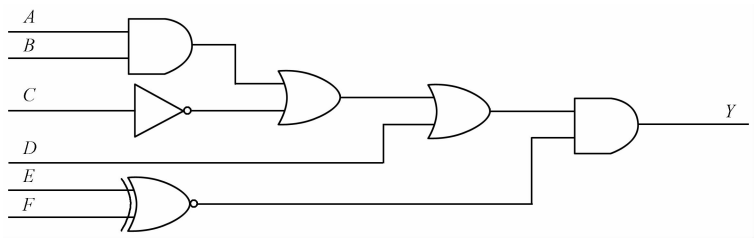


图 2.6.2 逻辑函数的电路图

2.7 Number System

2.7.1 Text

In this section we discuss number systems, and in particular we describe the various ways that numbers can be expressed in a binary representation, i. e. in a number system that uses two digits only: 0 and 1. The binary representation of information is generally utilized in computers and other digital systems because the basic elements from which such systems are built are of a binary nature^[1]. An understanding of the binary number representation is a prerequisite for most of the subsequent material.

Positional Notation

Representation of numbers in positional notation has received such universal acceptance that we tend to overlook its importance and the long historical evolution that brought us to this point^[2]. When we are given a sequence of decimal digits, e. g. 40795, we attach to each digit a multiplying factor, or weight that is some power of 10 and which depends on the position of the digit in the number, for example, $40795 = 4 \times 10^4 + 0 \times 10^3 + 7 \times 10^2 + 9 \times 10^1 + 5 \times 10^0$ ^[3].

Thus every decimal number of n digits is a sum of the weighted coefficients

$$N = N_r + n_r = [a_{n-1}(r)_{n-1} + \dots + a_0(r)^0 \cdot a_{-1}(r)^{-1} + \dots + a_{-m}(r)^{-m}]r = \sum a_i(r)^i \quad (2.7.1)$$

where a_i is the coefficient of the weight 10^i .

The positional notation is a short representation of Equation(2.7.1) in which the plus signs and the weights have been omitted. Thus

$$N_{10} = a_{n-1}a_{n-2} \dots a_1a_0 \quad (2.7.2)$$

Representation of numbers as shown in Equation(2.7.2) is barely ten centuries old. Early man used tally marks for counting. The Romans utilized a decimal system in which the position of the digit had some limited significance. The symbols used were I(1), V(5), X(10), L(50), C(100), CIC(1000), CCICC(10⁴), and CCCICCC(10⁵). Though an improvement over the tally system, arithmetic with Roman numbers was extremely cumbersome. Notice that the symbol for 0 is absent in this system. It appears that the Mesopotamians understood the concept of 0 and had a positional system using 60 symbols^[4]. This system was abandoned around 1700 B. C. although its influence is still witnessed today: 60 seconds in a minute, 60 minutes in one hour and 360 degrees in a circle. The decimal system with a positional notation was developed by the Hindus around the 5th century and was introduced to western civilization by the Arabs who also added the symbol 0^[5].

Referring back to Equation (2.7.2) we notice the general form of this positional representation. The base, or radix, 10 is only implied in that representation. We could have indeed used any other radix r . Equation (2.7.1) in its general form may thus be written

$$N_r = a_{n-1}(r)^{n-1} + a_{n-2}(r)^{n-2} + \cdots + a_1(r)^1 + a_0(r)^0 = \sum a_i(r)^i \tag{2.7.3}$$

Where r is the base in which the number is represented. Different symbols are required in a number system of radix r having a range of integers from 0 to $(r-1)$ as shown in Table 2.7.1 for $r = 2, 3, 8, 10, 12$ and 16 . Notice that for $r > 10$ we had to introduce additional symbols.

Table 2.7.1 Digits (symbols) used in several number systems.

Radix	Number system	Digital used in number system
2	Binary	0, 1
3	Ternary	0, 1, 2
8	Octal	0, 1, 2, 3, 4, 5, 6, 7
10	Decimal	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
12	Duodecimal	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B
16	Hexadecimal	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

The above discussion can be easily extended to express fractional numbers. In the decimal system we represent a fractional number by the digits that are placed to the right of the decimal point. This point of division between the integer and fractional parts of a number in any base is called the radix point. For example, the octal number (base 8) 0.1472 has the value $1 \times 8^{-1} + 4 \times 8^{-2} + 7 \times 8^{-3} + 2 \times 8^{-4}$. In general, the value of any fractional number n_r to the base r having m digits after the radix point is

$$n_r = a_{-1}(r)^{-1} + a_{-2}(r)^{-2} + \cdots + a_{-(m-1)}(r)^{-(m-1)} + a_{-m}(r)^{-m} = \sum a_i(r)^i \tag{2.7.4}$$

Combining Equations (2.7.3) and (2.7.4) we obtain a general expression for any integer, fractional or mixed number of $(n + m)$ digits in base r

$$N = N_r + n_r = [a_{n-1}(r)^{n-1} + \cdots + a_0(r)^0 + a_{-1}(r)^{-1} + \cdots + a_{-m}(r)^{-m}]r = \sum a_i(r)^i \tag{2.7.5}$$

Note the radix point between the r^0 -th and the r^{-1} -th terms above.

A subscript after a number will be used to indicate the radix of that number. For example, 143_{10} is a number in base 10, while 143_{16} and 143_8 are numbers in base 16 and 8, respectively. All three numbers have different numerical values, each given by Equation (2.7.5).

The Binary Number System

In the binary number system $r = 2$, thus the two digits used to represent any

number are 0 and 1. The digit 2 does not exist in this system. However, the numerical equivalent of 2 represented in the binary system can be obtained by application of simple arithmetic rules, remembering that a number greater than 1 generates a carry. Thus $1+1$ equals 0, carry 1, i. e. $1+1 = 10_2$; $10_2+10_2 = 100_2 = 4_{10}$; etc. Table 2.7.2 shows decimal numbers 0 through 20_{10} in column 2, 3, and 4, followed by selected numbers up to 1000_{10} .

The common use of binary and decimal numbers in digital systems requires procedures for converting a number given in one base to an equivalent number in another base. These will be discussed in detail in the next section. In what follows here, a few simple binary-to-decimal and decimal-to-binary conversion methods are described.

Binary-to-Decimal Conversion. Two binary-to-decimal conversion methods will be given; the first one of these is based on Equations (2.7.3)~(2.7.5).

Table 2.7.2 Representation of selected numbers in various number systems.

Decimal	Binary	Octal	Hexadecimal
0	0	0	0
1	1	1	1
2	10	2	2
3	11	3	3
4	100	4	4
5	101	5	5
6	110	6	6
7	111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F
16	10000	20	10
17	10001	21	11
18	10010	22	12
19	10011	23	13
20	10100	24	14
32	100000	40	20
50	110010	62	32
60	111100	74	3C
64	1000000	100	40
100	1100100	144	64
255	11111111	377	FF
1000	1111101000	1750	3E8

EXAMPLE 2.7.1 Convert to base ten the binary integer number $N_2 = 110101_2$.

From Equation(2.7.3), this number represents the sum

$$\begin{aligned}
 N_2 &= 1 \times (2)^5 + 1 \times (2)^4 + 0 \times (2)^3 + 1 \times (2)^2 + 0 \times (2)^1 + 1 \times (2)^0 \\
 &= (32 + 16 + 4 + 1)_{10} = 53_{10}
 \end{aligned}$$

Another binary-to-decimal conversion technique is the so-called double-and-add method: starting with the most significant binary digits (bit), double the 1 present in that bit and proceed to the next bit ^[6]. If the next bit is 0, double the number obtained in the previous step; if it is 1, add a 1 after the doubling. Continue until the least significant bit has been processed.

EXAMPLE 2.7.2 Convert to base 10 the binary number 101101_2 . The process is shown in Figure 2.7.1, yielding 45_{10} .

Decimal-to-Binary Conversion. One method for decimal-to-binary conversion of small numbers is based on recognition of the sum of powers of 2 contained in the decimal number ^[7]. The largest power of 2 is recognized first and subtracted from the decimal number. The same method is repeated on the difference obtained until the least significant bit has been processed.

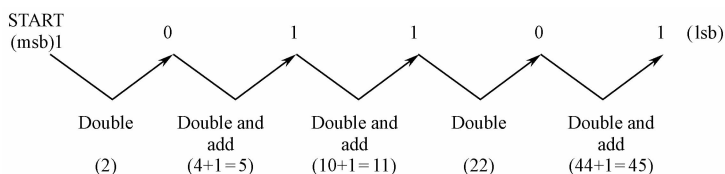


Figure 2.7.1 Binary-to-decimal conversion using the double-and-add method.

EXAMPLE 2.7.3 Convert to binary the decimal number 81_{10} .

The highest power of 2 contained in 81_{10} is $2^6 = 64_{10}$. Thus $(81 - 64)_{10} = 17_{10}$. The highest power of 2 contained in 17_{10} is $2^4 = 16_{10}$. Thus $(17 - 16)_{10} = 1$. The remainder represents the least significant bit, $1 = 2^0$. Thus $81_{10} = 1 \times 2^6 + 1 \times 2^4 + 1 \times 2^0$, and in positional notation $81_{10} = 1010001_2$.

The method of Example 2.7.3 is cumbersome for large numbers. Another algorithm is given here: The decimal number is divided by 2; the first remainder, 0 or 1, represents the least significant bit (lsb); the result is further divided by 2, the remainder representing the next bit (weight = 2^1), etc.

EXAMPLE 2.7.4 Convert to binary the decimal number 100_{10} .

	Remainder	Binary Weight
$100 \div 2 = 50$	0	2^0 (lsb)
$50 \div 2 = 25$	0	2^1
$25 \div 2 = 12$	1	2^2
$12 \div 2 = 6$	0	2^3
$6 \div 2 = 3$	0	2^4
$3 \div 2 = 1$	1	2^5
$1 \div 2 = 0$	1	2^6 (msb)

The answer is obtained by reading the remainder column from bottom to top:
 $100_{10} = 1100100_2$.

2.7.2 Specialized English Words

binary representation	二进制表示法	term	项
carry	进位(位)	coefficient	系数
digit	数字, 数位, 数码	subscript	下标
octa	八进制的	short representation	缩写(形式)
digital arithmetic circuits	数字算术 (运算)电路	binary-to-decimal conversion	二-十进 进制转换
hexadecimal	十六进制的	plus sign	加号
positional notation	位置计数法	double-and add method	“乘2加1”法
fractional numbers	分数	tally marks	筹码
representation	表示, 表达	remainder	余数
decimal point	十进制小数点	radix	基
multiplying factor	乘因子	least significant bit	最低有效位
radix point	小数点	equation	等式, 方程
weight	权	most significant bit	最高有效位
duodemical	十二进制的	binary number	二进制数
power	幂, 乘方	difference	差(数)

2.7.3 Notes

[1]The binary representation of information is generally utilized in computers and other digital systems because the basic elements from which such systems are built are of a binary nature. 不难看出,这是一个典型的主从复合句。“because”引起的是原因状语从句。从句的主语“the basic elements”又带有一个定语从句“from which such systems are built”。全句可译为“用二进制表示信息广泛用于计算机和其他数字系统,因为构建这些系统的基本元素具有二进制的本质。”

[2]Representation of numbers in positional notation has received such universal acceptance that we tend to overlook its importance and the long historical evolution that brought us to this point. 句中第一个“that”引导状语从句,表示结果;第二个“that”引导定语从句,修饰“evolution”。全句可译为“用位置计数法表示数得到如此广泛的认同,以至于人们往往对它熟视无睹,忽视了它的重大意义,以及人类走到这一步所经历的漫长历史发展过程。”

[3]When we are given a sequence of decimal digits, e. g. 40795, we attach to each digit a multiplying factor, or weight that is some power of 10 and which depends on the position of the digit in the number, for example, $40795 = 4 \times 10^4 + 0 \times 10^3 + 7 \times 10^2 + 9$

$\times 10^1 + 5 \times 10^0$ 。句中的“a sequence of decimal digits”指一串十进制数位组成的数,即一个十进制数。“some power of 10”为“10 的某次幂”。全句可译为“当我们得到一个十进制数,例如 40 795 时,我们给每一位数码加上一个乘因子,或者说‘权’,它是 10 的某次幂,幂的大小取决于该数码在整个数字中所处的位置,例如 $40\,795 = 4 \times 10^4 + 0 \times 10^3 + 7 \times 10^2 + 9 \times 10^1 + 5 \times 10^0$ 。”

[4] It appears that the Mesopotamians understood the concept of 0 and had a positional system using 60 symbols. 句中的“Mesopotamians”指美索不达米亚人,即古伊拉克人。“Positional system”指一种位置法计数系统。全句可译为“美索不达米亚人似乎懂得了‘0’的概念,并有了一种采用 60 个符号的位置记数系统。”

[5] The decimal system with a positional notation was developed by the Hindus around the 5th century and was introduced to western civilization by the Arabs who also added the symbol 0. 句中的“decimal system with a positional notation”指“十进制的位置计数法”。“Hindu”为印度人,“Arabs”为阿拉伯人。全句可译为“十进制位置计数法是印度人在大约 5 世纪发明的,并由阿拉伯人传给西方文明社会,阿拉伯人还增加了符号‘0’。”

[6] Another binary-to-decimal conversion technique is the so-called double-and-add method: starting with the most significant binary digits (bit), double the ‘1’ present in that bit and proceed to the next bit. 句中的“double-and-add”意思是“(先)乘 2(然后)再加”。根据操作方法可译为“乘 2 加 1”法。“Double”为动词,“the 1”为其宾语,而“present in that bit”为“the one”的定语。全句可译为“另一种二-十转换法被称为‘乘 2 加 1’法:从二进制最高位开始,将该位的 1 乘以 2,然后再处理下一位。”

[7] One method for decimal-to-binary conversion of small numbers is based on recognition of the sum of powers of 2 contained in the decimal number. 句中的“recognition”意为“认可、找出”。全句可译为“一种小数字的十-二转换方法是靠找出十进制数中所包含的全部 2 的幂的和。”

2.7.4 Reference Translation

数 字 系 统

在本节中,我们讨论数字系统,重点讨论二进制表示法,即使用“0”和“1”两个数码的数字系统中的数的各种表达方式。用二进制表示信息广泛用于计算机和其他数字系统,因为构建这些系统的基本元素具有二进制的本质。了解二进制表示法是学习后继内容的前提。

位置计数法

用位置计数法表示数得到如此广泛的认同,以至于人们往往对它熟视无睹,忽视了它的重大意义,以及人类走到这一步所经历的漫长历史发展过程。当我们得到一个十进制数,例如 40 795 时,我们给每一位数码加上一个乘因子,或者说“权”,它是 10 的某次幂,幂的大小取决于该数码在整个数字中所处的位置。例如 $40\,795 = 4 \times 10^4 + 0 \times 10^3 + 7 \times$

$$10^2 + 9 \times 10^1 + 5 \times 10^0。$$

所以,每一个 n 位的十进制数是其加权的系数之和:

$$N = N_r + n_r = [a_{n-1}(r)^{n-1} + \cdots + a_0(r)^0 \cdot a_{-1}(r)^{-1} + \cdots + a_{-m}(r)^{-m}]r = \sum a_i(r)^i \tag{2.7.1}$$

其中 a_i 为权 10^i 的系数。

位置计数法是式(2.7.1)省略了加号和权值的缩写。因而

$$N_{10} = a_{n-1}a_{n-2}\cdots a_1a_0 \tag{2.7.2}$$

用式(2.7.2)所示的方法表示数的历史不过十个世纪。早期人类使用筹码计数。罗马人使用了一种十进制,其中数码的位置具有某些有限的意义。所用的符号有 I(1)、V(5)、X(10)、L(50)、C(100)、CIC(1000)、CCICC(10⁴)及 CCCICCC(10⁵)等。尽管比起筹码计数是一个进步,但用罗马数进行算术运算极为麻烦。请注意,在这个系统中表示“0”的符号是没有的。美索不达米亚人似乎懂得了“0”的概念,并有了一种采用 60 个符号的位置计数系统。这套系统在公元前 1700 年左右被摒弃了,但它的影响至今仍然可见:1 分有 60 秒,1 小时有 60 分,一周有 360 度。十进制位置计数法是印度人在大约 5 世纪时发明的,并由阿拉伯人传给西方文明社会,阿拉伯人还增加了符号“0”。

让我们再观察式(2.7.2)这个位置计数法的一般表达式。它的基,恰好是 10 这个数,其实完全可以采用任何其他的基。所以式(2.7.1)的一般形式可以写为

$$N_r = a_{n-1}(r)^{n-1} + a_{n-2}(r)^{n-2} + \cdots + a_1(r)^1 + a_0(r)^0 = \sum a_i(r)^i \tag{2.7.3}$$

其中 r 是所表示的数的基。在基为 r 的数字系统中,不同的 r 需要有一组从 0 到 $(r-1)$ 的不同的整数符号。表 2.7.1 给出的是 $r = 2、3、8、10、12$ 和 16 时的情况。注意当 $r > 10$ 时必须引入新的符号。

表 2.7.1 几种数字系统所用的数码(符号)

基	数字系统	数字系统所用数码
2	二进制	0,1
3	三进制	0,1,2
8	八进制	0,1,2,3,4,5,6,7
10	十进制	0,1,2,3,4,5,6,7,8,9
12	十二进制	0,1,2,3,4,5,6,7,8,9,A,B
16	十六进制	0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F

上述讨论很容易推广到小数。在十进制中,我们用放在小数点右边的数字来表示小数。这个在一个数的整数部分和小数部分之间用来进行分隔的圆点都被称为基点(小数点)。例如,八进制数(基 8)0.1472 的值为 $1 \times 8^{-1} + 4 \times 8^{-2} + 7 \times 8^{-3} + 2 \times 8^{-4}$ 。一般而言,任何以 r 为基,小数点后数位为 m 的分数 n_r 的值为

$$n_r = a_{-1}(r)^{-1} + a_{-2}(r)^{-2} + \cdots + a_{-(m-1)}(r)^{-(m-1)} + a_{-m}(r)^{-m} = \sum a_i(r)^i \tag{2.7.4}$$

将式(2.7.3)和式(2.7.4)结合起来,可以得到基为 r ,具有 $(n+m)$ 位的任何整数、分

数以及混合数的一般表达式：

$$N = N_r + n_r = [a_{n-1}(r)^{n-1} + \cdots + a_0(r)^0 \cdot a_{-1}(r)^{-1} + \cdots + a_{-m}(r)^{-m}]r = \sum a_i(r)^i$$

(2.7.5)

注意上式中的第 r^0 项和第 r^{-1} 项之间的小数点。

用数字后的下标表示该数的基。例如,143₁₀ 表示一个以 10 为基的数。而 143₁₆ 和 143₈ 的基分别为 16 和 8。三个数的数值各不相同,式(2.7.5)可以给出每个值。

表 2.7.2 不同数字系统的表示举例

十进制	二进制	八进制	十六进制
0	0	0	0
1	1	1	1
2	10	2	2
3	11	3	3
4	100	4	4
5	101	5	5
6	110	6	6
7	111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F
16	10000	20	10
17	10001	21	11
18	10010	22	12
19	10011	23	13
20	10100	24	14
32	100000	40	20
50	110010	62	32
60	111100	74	3C
64	1000000	100	40
100	1100100	144	64
255	11111111	377	FF
1000	1111101000	1750	3E8

二进制数字系统

在二进制数字系统中, $r = 2$,所以用于表示任何数字的两个数码为“0”和“1”。数码 2 在这个系统中是不存在的。但 2 的等值表达在二进制中可以通过简单的算术运算规则得到。记住,大于 1 的数会产生一个进位。所以 1+1 等于 0,进位 1,即 $1+1 = 10_2$; $10_2+10_2 = 100_2 = 4_{10}$,等等。表 2.7.2 第一列给出 1 至 20₁₀ 及若干个其他 1000₁₀ 以内的十进制数。第二、三、四列则是对应的二进制、八进制和十六进制数。

在数字系统中使用二进制和十进制数时,需要将用某个基表示的数等值转换成用另一个基表示的数。这个问题将在下一节中详细讨论。下面介绍几个简单的二-十转换和

十-二转换方法。

二十转换。下面给出两个二十转换的方法。第一个方法是以式(2.7.3)~(2.7.5)为依据的。

例 2.7.1 将二进制整数 $N_2 = 110101_2$ 转换为以十为基的数。

根据式(2.7.3),该数表示和数如下:

$$\begin{aligned} N_2 &= 1 \times (2)^5 + 1 \times (2)^4 + 0 \times (2)^3 + 1 \times (2)^2 + 0 \times (2)^1 + 1 \times (2)^0 \\ &= (32 + 16 + 4 + 1)_{10} = 53_{10} \end{aligned}$$

另一种二十转换方法被称为“乘 2 加 1”法:从二进制最高位开始,将该位的 1 乘以 2,然后再处理下一位。如果下一位是 0,则将上一位得到的数再乘 2;如果下一位是 1,则先加 1 再乘 2。如此进行直到最低位处理完为止。

例 2.7.2 将二进制数 101001_2 转换为十进制。图 2.7.1 给出了转换过程,得到 45_{10} 。

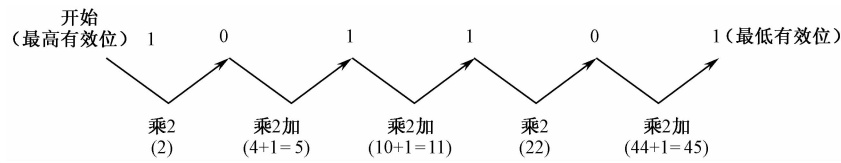


图 2.7.1 用“乘 2 加 1”法进行二-十转换

十-二转换。对于数值不大的数,一种小数字的十-二转换方法是靠找出十进制数中所包含的全部 2 的幂的和。先找 2 的最大的幂,然后将它从该数中减去。对所得的差重复同样的做法,直到最低位处理完毕。

例 2.7.3 将十进制数 81_{10} 转换为二进制数。

81_{10} 中所含的最大的 2 的幂是 $2^6 = 64_{10}$ 。所以 $(81 - 64)_{10} = 17_{10}$ 。 17_{10} 中所含最大的 2 的幂是 $2^4 = 16_{10}$ 。所以 $(17 - 16)_{10} = 1$ 。最后剩下的是最低位数 $1 = 2^0$ 。所以 $81_{10} = 1 \times 2^6 + 1 \times 2^4 + 1 \times 2^0$,用位置计数法表示即 $81_{10} = 1010001_2$ 。

例 3 所给的方法对于大数就十分麻烦。下面给出另一种方法。将十进制数除以 2;第一个余数为 0 或 1,代表最低位;将商再除以 2,余数代表下一位(权 $= 2^1$),等等。

例 4 将十进制数 100_{10} 转换为二进制数。

十进制数 100_{10} 转换为二进制数

	余 数	二进制权
$100 \div 2 = 50$	0	2^0 (最低有效位)
$50 \div 2 = 25$	0	2^1
$25 \div 2 = 12$	1	2^2
$12 \div 2 = 6$	0	2^3
$6 \div 2 = 3$	0	2^4
$3 \div 2 = 1$	1	2^5
$1 \div 2 = 0$	1	2^6 (最高有效位)

从余数列最下行向上读,便得到答案: $100_{10} = 1100100_2$ 。

2.7.5 Reading Materials

Basic Digital Concepts

By converting continuous analog signals into a finite number of discrete states, a process called digitization(数字化), then to the extent that the states are sufficiently well separated so that noise does create errors, the resulting digital signals allow the following (slightly idealized):

- storage over arbitrary periods of time.
- flawless retrieval(无错检索) and reproduction of the stored information.
- flawless transmission(无错传送) of the information.

Some information is intrinsically digital, so it is natural to process and manipulate it using purely digital techniques. Examples are numbers and words.

The drawback(缺点) to digitization is that a single analog signal (e.g. a voltage which is a function of time, like a stereo signal) needs many discrete states, or bits, in order to give a satisfactory reproduction. For example, it requires a minimum of 10 bits to determine a voltage at any given time to an accuracy of $\approx 0.1\%$. For transmission, one now requires 10 lines instead of the one original analog line.

The explosion in digital techniques and technology has been made possible by the incredible increase in the density of digital circuitry, its robust performance(性能稳定), its relatively low cost, and its speed. The requirement of using many bits in reproduction is no longer an issue: The more the better.

This circuitry is based upon the transistor, which can be operated as a switch with two states. Hence, the digital information is intrinsically(本质上) binary. So in practice, the terms digital and binary are used interchangeably. In the following sections we summarize some conventions for defining the binary states and for doing binary arithmetic.

2.8 Flip-Flops and Latches

2.8.1 Text

Logic alone does not a system make^[1]. Boolean equations provide the means to transform a set of inputs into deterministic results. However, these equations have no ability to store the results of previous calculations upon which new calculations can be made^[2]. The preceding adder logic continually recalculates the sum of two inputs. If either input is removed from the circuit, the sum disappears as well. A series of numbers that arrive one at a time cannot be summed, because the adder has no means of storing a running total. Digital systems operate by maintaining state to advance through

sequential steps in an algorithm. State is the system’s ability to keep a record of its progress in a particular sequence of operations. A system’s state can be as simple as a counter or an accumulated sum.

State-full logic elements called flip-flops are able to indefinitely hold a specific state (0 or 1) until a new state is explicitly loaded into them. Flip-flops load a new state when triggered by the transition of an input clock. A clock is a repetitive binary signal with a defined period that is composed of 0 and 1 phases as shown in Figure 2. 8. 1. In addition to a defined period, a clock also has a certain duty cycle, the ratio of the duration of its 0 and 1 phases to the overall period. An ideal clock has a 50/50 duty cycle, indicating that its period is divided evenly between the two states^[3]. Clocks regulate the operation of a digital system by allowing time for new results to be calculated by logic gates and then capturing the results in flip-flops.

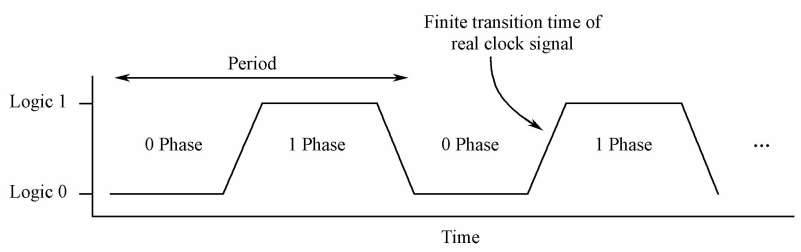


Figure 2. 8. 1 Digital clock signal.

There are several types of flip-flops, but the most common type in use today is the D flip-flop. Other types of flip-flops include RS and JK, but this discussion is restricted to D flip-flops because of their standardized usage. A D flip-flop is often called a flop for short. A basic rising-edge triggered flop has two inputs and one output as shown in Figure 2. 8. 2(a). By convention, the input to a flop is labeled D, the output is labeled Q, and the clock is represented graphically by a triangle. When the clock transitions from 0 to 1, the state at the D input is propagated to the Q output and stored until the next rising edge. State-full logic is often described through the use of a timing diagram, a drawing of logic state versus time. Figure 2. 8. 2(b) shows a basic flop timing diagram in which the clock’s rising edge triggers a change in the flop’s state. Prior to the rising edge, the flop has its initial state, Q_0 , and an arbitrary 0 or 1 input is applied as D_0 . The rising edge loads D_0 into the flop, which is reflected at the output. Once triggered, the flop’s input can change without affecting the output until the next rising edge. Therefore, the input is labeled as “don’t care,” or “xxx” following the clock’s rising edge.

Rising-edge flops are the norm, although some flops are falling-edge triggered. A

falling-edge triggered flop is indicated by placing an inversion bubble at the clock input as shown in Figure 2. 8. 3. Operation is the same, with the exception that the polarity of the clock is inverted.

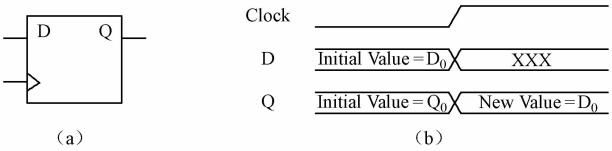


Figure 2. 8. 2 Rising-edge triggered flop.

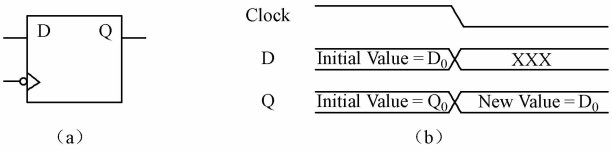


Figure 2. 8. 3 Falling-edge triggered flop.

There are several common feature enhancements to the basic flop, including clock-enable, set, and clear inputs and a complementary output. Clock enable is used as a triggering qualifier each time a rising clock edge is detected. The D input is loaded only if clock enable is set to its active state. Inputs in general are defined by device manufacturers to be either active-low or active-high. An active-low signal is effective when set to 0, and an active-high signal is effective when set to 1. Signals are assumed to be active-high unless otherwise indicated. Active-low inputs are commonly indicated by the same inversion bubble used to indicate a falling-edge clock. When a signal is driven to its active state, it is said to be asserted. A signal is de-asserted when driven to its inactive state. Set and clear inputs explicitly force a flop to a 1 or 0 state, respectively. Such inputs are often used to initialize a digital system to a known state when it is first turned on. Otherwise, the flop powers up in a random state, which can cause problems for certain logic. Set and clear inputs can be either synchronous or asynchronous. Synchronous inputs take effect only on the rising clock edge, while asynchronous inputs take effect immediately upon being asserted. A complementary output is simply an inverted copy of the main output.

A truth table for a flop enhanced with the features just discussed is shown in Table 2. 8. 1. The truth table assumes a synchronous, active-high clock enable (EN) and synchronous, active-low set and clear inputs. The rising edge of the clock is indicated by the \uparrow symbol. When the clock is at either static value, the outputs of the flop remain in their existing states. When the clock rises, the D, EN, $\overline{\text{CLR}}$ and $\overline{\text{SET}}$

inputs are sampled and acted on accordingly. As a general rule, conflicting information such as asserting $\overline{\text{CLR}}$ and $\overline{\text{SET}}$ at the same time should be avoided, because unknown results may arise. The exact behavior in this case depends on the specific flop implementation and may vary by manufacturer.

Table 2. 8. 1 Enhanced Flop Truth Table

Clock	D	EN	$\overline{\text{CLR}}$	$\overline{\text{SET}}$	Q	$\overline{\text{Q}}$
0	X	X	X	X	Q _{static}	$\overline{\text{Q}}_{\text{static}}$
↑	0	0	1	1	Q _{static}	$\overline{\text{Q}}_{\text{static}}$
↑	0	1	1	1	0	1
↑	1	1	1	1	1	0
↑	X	X	0	1	0	1
↑	X	X	1	0	1	0
↑	X	X	0	0	?	?
1	X	X	X	X	Q _{static}	$\overline{\text{Q}}_{\text{static}}$

A basic application of flops is a binary ripple counter. Multiple flops can be cascaded as shown in Figure 2. 8. 4 such that each complementary output is fed back to that flop’s input and also used to clock the next flop. The current count value is represented by the noninverted flop outputs with the first flop representing the LSB. A three-bit counter is shown with an active-low reset input so that the counter can be cleared to begin at zero. The counter circuit diagram uses the standard convention of showing electrical connectivity between intersecting wires by means of a junction dot. Wires that cross without a dot at their intersection are not electrically connected.

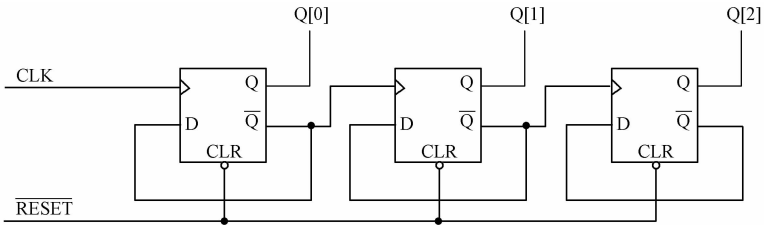


Figure 2. 8. 4 Three-bit ripple counter.

The ripple counter’s operation is illustrated in Figure 2. 8. 5. Each bit starts out at zero if $\overline{\text{RESET}}$ is asserted^[4]. Counting begins on the first rising edge of CLK following the de-assertion of $\overline{\text{RESET}}$. The LSB, Q[0], increments from 0 to 1, because its D input is driven by the complementary output, which is 1. The complementary output transitions to 0, which does not trigger the Q[1] rising-edge flop, but IT does set up the conditions for a trigger after the next CLK rising edge. When CLK rises again, Q[0] transitions back to 0, and Q[0] transitions to 1, forming a rising edge to trigger Q[1], which loads a 1. This sequence continues until the count value reaches 7,

at which point the counter rolls over to zero, and the sequence begins again^[5].

An undesirable characteristic of the ripple counter is that it takes longer for a new count value to stabilize as the number of bits in the counter increases. Because each flop’s output clocks the next flop in the sequence, it can take some time for all flops to be updated following the CLK rising edge^[6]. Slow systems may not find this burdensome, but the added ripple delay is unacceptable in most highspeed applications. Ways around this problem will be discussed shortly.

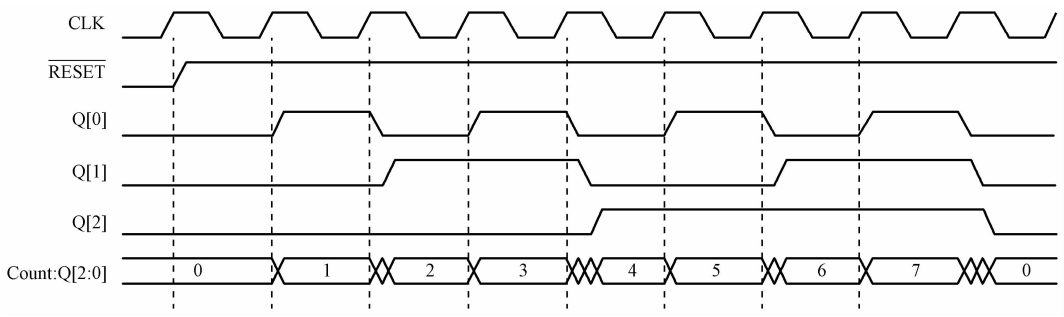


Figure 2. 8. 5 Ripple counter timing diagram.

2. 8. 2 Specialized English Words

deterministic	确定性的	truth table	真值表
flip-flops	触发器	ripple counter	行波计数器
duty cycle	占空比	asserted	确定的
rising-edge	上升沿	undesirable	不希望的,不受欢迎的
triggered	触发	characteristic	特性
triangle	三角形	stabilize	稳定
falling-edge	下降沿	burdensome	难以承受的
enhancements	增强	powers up	上电
active-low	低(电平)有效		

2. 8. 3 Notes

- [1]Logic alone does not a system make. 这是一个倒装句。正常语序应为“Logic alone does not make a system.”。全句可译为“逻辑不可能单独地构成一个完整的系统。”
- [2] However, these equations have no ability to store the results of previous calculations upon which new calculations can be made. 本句是一个含有定语从句的主从复合句,“upon which”引起的定语从句修饰“calculations”。全句可译为“然而,逻辑方程在进行新的计算时没有存储过去计算结果的能力。”

[3]An ideal clock has a 50/50 duty cycle, indicating that its period is divided evenly between the two states. 句中“indicating… states”为状语,修饰主句谓语动词“has”,“indicating”为现在分词,“that”引起的是动名词从句,作为“indicating”的宾语。全句可译为“一个理想时钟具有 50/50 的占空比,即整个时钟周期平均地分配给两个状态。”

[4]Each bit starts out at zero if $\overline{\text{RESET}}$ is asserted. 句中“if $\overline{\text{RESET}}$ is asserted”为条件状语从句,用做主句谓语动词“starts out”的条件。全句可译为“如果复位信号 $\overline{\text{RESET}}$ 有效,每一位则从 0 开始计数。”

[5]This sequence continues until the count value reaches 7, at which point the counter rolls over to zero, and the sequence begins again. 本句主句为“This sequence continues”,“until”为连接词,其后的“the count value reaches 7”为它所引导的从句,修饰主句中“continues”。句中“at which”以下为定语从句,修饰“7”,从句中“rolls”和“begins”为并列动词。全句可译为“这个顺序一直持续到值为 7,然后计数器循环变为 0,接着重新开始计数。”

[6]Because each flop’s output clocks the next flop in the sequence, it can take some time for all flops to be updated following the CLK rising edge. 本句中“it can take … edge.”为主句,“Because… sequence”为其原因从句。注意主句中“following the CLK rising edge”,为现在分词短语做状语,修饰“be updated”。全句可译为“这是因为在这个过程中,每个触发器的输出都是下一个触发器的时钟输入。随着 CLK 上升沿的出现,所有触发器都更新就需要一定的时间。”

2.8.4 Reference Translation

触发器和锁存器

逻辑不可能单独地构成一个完整的系统。布尔方程只是提供了一种将一组输入转换为一个确定结果的方式。然而,逻辑方程在进行新的计算时没有存储过去计算结果的能力。如加法逻辑连续计算两个输入的和,如果取消加法电路的任何一个输入,计算和也会随之消失。即使同一时刻有多个数据要加,由于加法器没有存储运行结果的功能,我们也不能得到最终的加法和。数字系统正是通过维持某种状态的方式来完成具有连续步骤的算法。状态是一种系统功能,它记录着连续操作算法的步骤。系统状态可以由简单的计数器或累加器实现。

触发器是一种状态存储单元,它能够维持一个状态直到需要载入一个新的状态。触发器的状态变化是靠输入时钟的变化触发的。时钟信号反复产生 0 和 1 两种信号,具有固定的时钟周期,如图 2.8.1 所示。除了固定周期外,时钟信号还有确定的占空比,规定了 0 相位和 1 相位的持续时间占整个周期的比例。一个理想时钟具有 50/50 的占空比,即整个时钟周期平均地分配给两个状态。在时钟信号控制下数字系统操作变成:逻辑门计算新的结果,触发器存储执行结果。

触发器有许多种,当今最常用的是 D 触发器,其他还包括 RS 触发器和 JK 触发器。

由于数字系统中广泛使用的是 D 触发器,因此我们讨论的重点仅限于 D 触发器。也将 D 触发器简称为触发器。一个基本的上升沿触发的触发器包括两个输入和一个输出,如图 2.8.2(a) 所示。按照习惯,触发器的输入用 D 标识,输出用 Q 标识,时钟用一个三角形标识。当时钟信号从 0 变 1 时, Q 输出的状态同 D 输入相同,同时保存 Q 值,直到下一个上升沿的到来。完整的状态逻辑通常通过时序图,即逻辑状态沿随时间的变化来描述。图 2.8.2(b) 是一个靠上升沿触发状态翻转的 D 型触发器的基本时序图。在上升沿来临之前,触发器的初始状态为 Q_0 ,同时 D_0 可以任意为 0 或 1。上升沿将 D_0 载入触发器中,并出现在输出端。一旦触发,直到下一个上升沿的来临之前触发器的输入的变化将不会影响触发器的输出。因此在图中紧接着上升沿的输入段标记为“无关”或“XXX”。

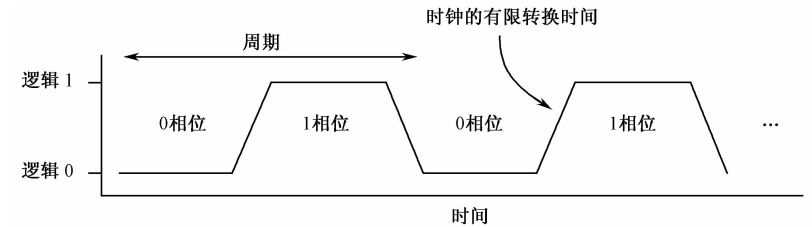


图 2.8.1 数字时钟信号

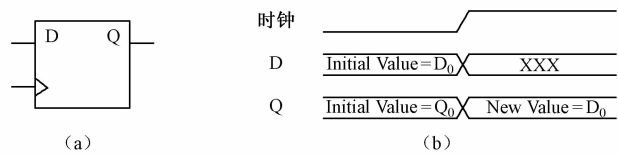


图 2.8.2 上升沿触发器

规范的触发器都是上升沿触发,但也有些是下降沿触发的。下降沿触发的触发器在时钟输入处用一个代表反相的圆圈来标识,如图 2.8.3 所示。除了时钟反相外,操作是一样的。

在基本触发器的基础上常有一些性能改进,如时钟控制、置位、复位以及互补输出等。时钟控制信号的作用像是一个触发监测器,对每一个时钟上升沿进行检测。只有当时钟控制信号有效时,D 端输入信号才能被有效地载入触发器。信号是低电平有效还是高电平有效,通常是由器件制造商规定的。低电平有效表示信号为 0 时有效,高电平有效则表示信号为 1 时有效。除非特别指出,我们都假定为高电平有效。低电平有效的输入常常也像用表示下降沿时钟的那个代表反相的圆圈来标识。

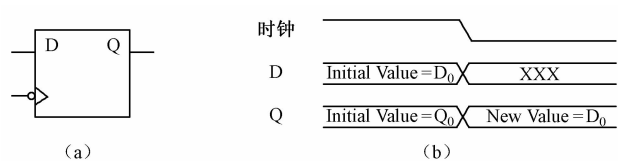


图 2.8.3 下降沿触发器

当信号被置于有效状态时,我们称该触发器被置位了,当信号被置于无效状态时,则称为被复位了。通过置位和复位可以分别将触发器的状态强制成 1 或 0。这样的输入常常用数字系统初始化,使系统第一次启动时触发器处于确定状态,否则触发器的状态将是随机的,这对某些逻辑而言将会发生问题。置位和复位可以是同步的也可以是异步的。同步输入仅在时钟上升沿时起作用,而异步输入则在信号一有效立即起作用。互补输出仅是主输出的反转。

具有刚才所讨论的增强功能的触发器的真值表见表 2.8.1。真值表假定该触发器具有同步高电平有效(EN)的时钟控制、同步低电平有效的置位和复位功能及互补输出。时钟上升沿用标号 \uparrow 表示。时钟不发生变化时,输出也保持不变。当时钟信号上升时, D、EN、 $\overline{\text{CLR}}$ 以及 $\overline{\text{SET}}$ 输入被采样,并按所讨论的起作用。作为一般原则,互相冲突的信号,例如置位 $\overline{\text{CLR}}$ 和复位 $\overline{\text{SET}}$,应避免同时产生,否则有可能导致不确定的结果,这时触发器的确切表现取决于具体的器件,且会因制造商而异。

表 2.8.1 增强型触发器真值表

Clock	D	EN	$\overline{\text{CLR}}$	$\overline{\text{SET}}$	Q	$\overline{\text{Q}}$
0	X	X	X	X	Q_{static}	$\overline{Q_{\text{static}}}$
\uparrow	0	0	1	1	Q_{static}	$\overline{Q_{\text{static}}}$
\uparrow	0	1	1	1	0	1
\uparrow	1	1	1	1	1	0
\uparrow	X	X	0	1	0	1
\uparrow	X	X	1	0	1	0
\uparrow	X	X	0	0	?	?
1	X	X	X	X	Q_{static}	$\overline{Q_{\text{static}}}$

二进制行波计数器是触发器的一种基本应用。多个触发器如图 2.8.4 所示级联在一起,每个触发器的反相输出既反馈到本触发器作为输入,又用作下一个触发器的时钟信号。当前的计数值由各个触发器的正相输出组成,第一个触发器的输出为最低位。图 2.8.4 所示的是一个三位计数器,有一个低电平有效的复位输入信号保证计数器能够复位从零开始计数。该计数器电路图中,当有导线相交时,采用了用连接点表示的标准电路连接方式。导线相交而没有连接点时,表示没有电路连接。

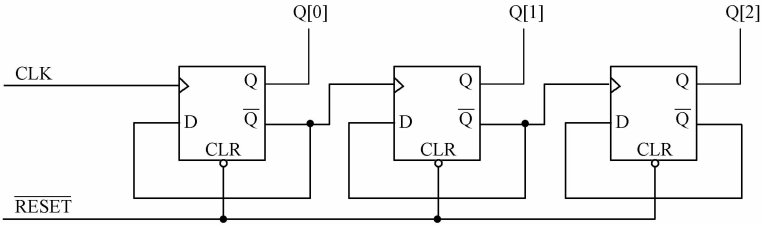


图 2.8.4 3 位行波计数器

行波计数器的操作过程如图 2.8.5 所示。如果复位信号 $\overline{\text{RESET}}$ 有效,每一位则从 0

开始计数。计数从 $\overline{\text{RESET}}$ 复位后的 CLK 的第一个上升沿开始。最低位 $Q[0]$ 从 0 增加到 1, 因为它的 D 输入端连接的是该触发器的反相输出, 值为 1。而触发器的反相输出变为 0, 不会触发上升沿触发的 $Q[1]$, 但是在下一个 CLK 上升沿到来之前已经建立了 $Q[1]$ 的触发条件。当第二个时钟上升沿到来时, $Q[0]$ 变回到 0, 而 $\overline{Q[0]}$ 变为 1, 从而形成一个上升沿触发 $Q[1]$, 输入一个 1。这个顺序一直持续到值为 7, 然后计数器循环变为 0, 接着重新开始计数。

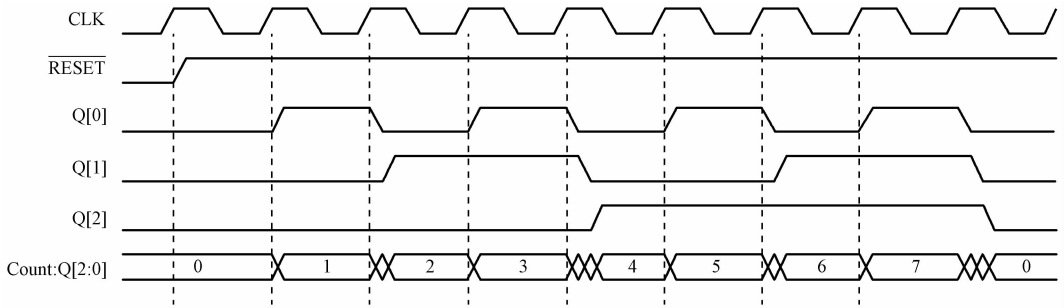


图 2.8.5 行波计数器时序图

行波计数器的不足之处就是当计数器的位数增加时需要较长时间才能使得新计数值稳定下来。这是因为在这个过程中, 每个触发器的输出都是下一个触发器的时钟输入。随着 CLK 上升沿的出现, 所有的触发器都更新就需要一定的时间。低速系统也许不觉得这是个问题, 但在大多数高速应用场合, 行波计数器的延迟是不可接受的。后面很快将讨论这个问题的解决方法。

2.9 Programmable Logic Device

2.9.1 Text

Programmable logic is the means by which a large segment of engineers implement their custom logic, whether that logic is a simple I/O port or a complex state machine^[1]. Most programmable logic is implemented with some type of HDL that frees the engineer from having to derive and minimize Boolean expressions each time a new logical relationship is designed. The advantages of programmable logic include rapid customization with relatively limited expense invested in tools and support.

The widespread availability of flexible programmable logic products has brought custom logic design capabilities to many individuals and smaller companies that would not otherwise have the financial and staffing resources to build a fully custom IC. These devices are available in a wide range of sizes, operating voltages, and speeds, which all but guarantees that a particular application can be closely matched with a relevant device. Selecting that device requires some research, because each manufacturer has a

slightly different specialty and range of products.

Among the most basic types of PLDs are Generic Array Logic (GAL) devices. GALs are enhanced variants of the older Programmable Array Logic (PAL) architecture that is now essentially obsolete. The term PAL is still widely used, but people are usually referring to GAL devices or other PLD variants when they use the term. PALs became obsolete, because GALs provide a superset of their functionality and can therefore perform all of the functions that PALs did. GALs are relatively small, inexpensive, easily available, and manufactured by a variety of vendors (e. g. , Cypress, Lattice, and Texas Instruments).

It can be shown through Boolean algebra that any logical expression can be represented as an arbitrarily complex sum of products^[2]. Therefore, by providing a programmable array of AND/OR gates, logic can be customized to fit a particular application^[3]. GAL devices provide an extensive programmable array of wide AND gates, as shown in Figure 2. 9. 1, into which all the device’s input terms are fed. Both true and inverted versions of each input are made available to each AND gate.

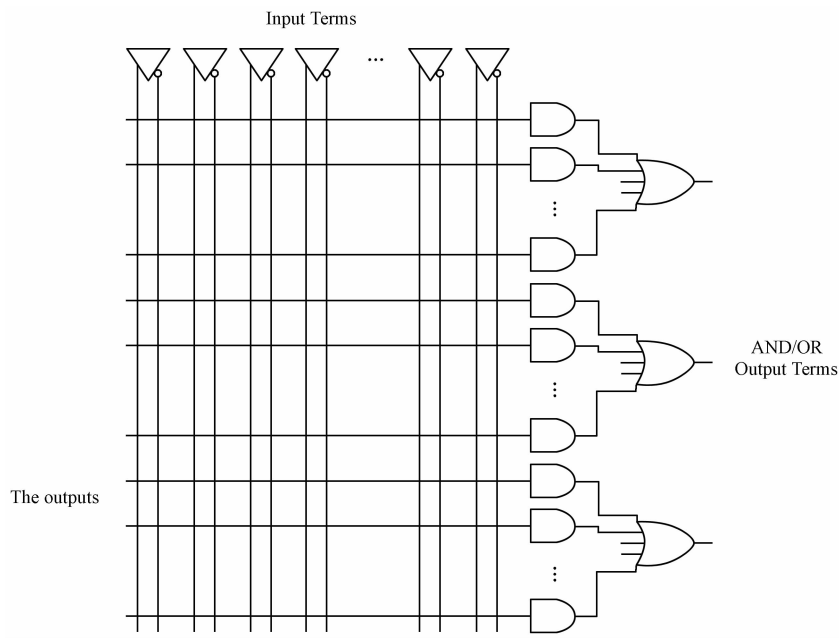


Figure 2. 9. 1 GAL/PAL AND/OR structure.

The outputs of groups of AND gates (products) feed into separate OR gates (sums) to generate user-defined Boolean expressions.

Each intersection of a horizontal AND gate line and a vertical input term is a programmable connection. In the early days of PLDs, these connections were made by

fuses that would literally have to be blown with a high voltage to configure the device. Fuse-based devices were not reprogrammable; once a microscopic fuse is blown, it cannot be restored. Today's devices typically rely on EEPROM technology and CMOS switches to enable nonvolatile reprogrammability. However, fuse-based terminology remains in use for historical reasons. The default configuration of a connection emulates an intact fuse, thereby connecting the input term to the AND gate input. When the connection is blown, or programmed, the AND input is disconnected from the input term and pulled high to effectively remove that input from the Boolean expression^[4]. Customization of a GAL's programmable AND gate is conceptually illustrated in Figure 2.9.2.

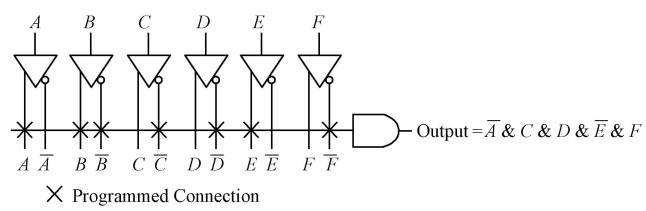


Figure 2.9.2 Programming AND input terms.

With full programmability of the AND array, the OR connections can be hard wired. Each GAL device feeds a differing number of AND terms into the OR gates. If one or more of these AND terms are not needed by a particular Boolean expression, those unneeded AND gates can be effectively disabled by forcing their outputs to 0. This is done by leaving an unneeded AND gate's inputs unprogrammed. Remember that inputs to the AND array are provided in both true and complement versions. When all AND connections are left intact, multiple expressions of the form $A \& \bar{A} = 0$ result, thereby forcing that gate's output to 0 and rendering it nonparticipatory in the OR function.

The majority of a GAL's logic customization is performed by programming the AND array. However, selecting flip-flops, OR/NOR polarities, and input/output configurations is performed by programming a configurable I/O and feedback structure called a macrocell^[5]. The basic concept behind a macrocell is to ultimately determine how the AND/OR Boolean expression is handled and how the macrocell's associated I/O pin operates. A schematic view of a GAL macrocell is shown in Figure 2.9.3, although some GALs may contain macrocells with slightly different capabilities. Multiplexers determine the polarity of the final OR/NOR term, regardless of whether the term is registered and whether the feedback signal is taken directly at the flop's output or at the pin. Configuring the macrocell's output enable determines how the pin behaves.

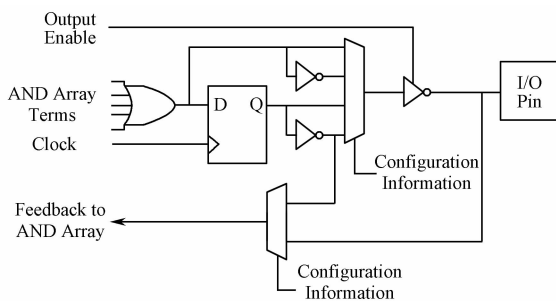


Figure 2. 9. 3 GAL macrocell.

There are two common GAL devices, the 16V8 and the 22V10, although other variants exist as well. They contain eight and ten macrocells each, respectively. The 16V8 provides up to 10 dedicated inputs that feed the AND array, whereas the 22V10 provides 12 dedicated inputs^[6]. One of the 22V10's dedicated inputs also

serves as a global clock for any flops that are enabled in the macrocells. Output enable logic in a 22V10 is evaluated independently for each macrocell via a dedicated AND term. The 16V8 is somewhat less flexible, because it cannot arbitrarily feed back all macrocell outputs depending on the device configuration. Additionally, when configured for registered mode where macrocell flops are usable, two dedicated input pins are lost to clock and output enable functions.

GALs are fairly low-density PLDs by modern standards, but their advantages of low cost and high speed are derived from their small size. Implementing logic in a GAL follows several basic steps. First, the logic is represented in either graphical schematic diagram or textual (HDL) form. This representation is converted into a netlist using a translation or synthesis tool. Finally, the netlist is fitted into the target device by mapping individual gate functions into the programmable AND array. Given the fixed AND/OR structure of a GAL, fitting software is designed to perform logic optimizations and translations to convert arbitrary Boolean expressions into sum-of-product expressions. The result of the fitting process is a programming image, also called a fuse map, that defines exactly which connections, or fuses, are to be programmed and which are to be left at their default state. The programming image also contains other information such as macrocell configuration and other device-specific attributes.

Modern PLD development software allows the back-end GAL synthesis and fitting process to proceed without manual intervention in most cases. The straightforward logic flow through the programmable AND array reduces the permutations of how a given Boolean expression can be implemented and results in very predictable logic fitting. An input signal propagates through the pin and pad structure directly into the AND array, passes through just two gates, and can then either feed a macrocell flop or drive directly out through an I/O pin. Logic elements within a GAL are close to each other as a result of the GAL's small size, which contributes to low internal propagation delays. These characteristics enable the GAL architecture to deliver very fast timing specifications,

because signals follow deterministic paths with low propagation delays.

GALs are a logic implementation technology with very predictable capabilities. If the desired logic cannot fit within the GAL, there may not be much that can be done without optimizing the algorithm or partitioning the design across multiple devices. If the logic fits but does not meet timing, the logic must be optimized, or a faster device must be found. Because of the GAL's basic fitting process and architecture, there isn't the same opportunity of tweaking the device as can be done with more complex PLDs. This should not be construed as a lack of flexibility on the part of the GAL. Rather, the GAL does what it says it does, and it is up to the engineer to properly apply the technology to solve the problem at hand. It is the simplicity of the GAL architecture that is its greatest strength.

Lattice Semiconductor's GAL22LV10D-4 device features a worst-case input-to-output combinatorial propagation delay of just 4 ns. This timing makes the part suitable for address decoding on fast microprocessor interfaces. The same 22V10 part features a 3 ns t_{co} and up to 250 MHz operation. The t_{co} specification is a pin-to-pin metric that includes the propagation delays of the clock through the input pin and the output signal through the output pin. Internally, the actual flop itself exhibits a faster t_{co} that becomes relevant for internal logic feedback paths. Maximum clock frequency specifications are an interesting aspect of all PLDs and some consideration. These specifications are best-case numbers usually obtained with minimal logic configurations. They may define the highest toggle rate of the device's flops, but synchronous timing analysis dictates that there is more to f_{MAX} than the flop's t_{SU} and t_{co} . Propagation delay of logic and connectivity between flops is of prime concern. The GAL architecture's deterministic and fast logic feedback paths reduces the added penalty of internal propagation delays. Lattice's GAL22LV10D features an internal clock-to-feedback delay of 2.5 ns, which is the combination of the actual flop's t_{co} plus the propagation delay of the signal back through the AND/OR array. This feedback delay, when combined with the flop's 3 ns t_{SU} , yields a practical f_{MAX} of 182 MHz when dealing with most normal synchronous logic that contains feedback paths (e. g. , a state machine).

2.9.2 Specialized English Words

programmable logic	可编程逻辑	PLDs	可编程逻辑器件
complex state machine	复杂状态机	Boolean algebra	布尔代数
HDL	硬件描述语言	AND gates	与门
IC	集成电路	OR gates	或门
devices	器件	Boolean expressions	布尔表达式

fuse-based	基于熔丝工艺的	textual (HDL) form	文本方式
flip-flops	触发器	programming image	编程映像
polarities	极性	fuse map	熔丝图
configurations	配置	synthesis	综合
macrocell	宏单元	characteristics	特性
I/O pin	输入/输出引脚	predictable capabilities	可预测容量
schematic	原理图	worst-case	最坏的情况
multiplexers	多路器	propagation delay	传播延迟
term	项	suitable for	适合……的
fitted into	适配到……	specification	规范
dealing with	处理	pin-to-pin	引脚到引脚
synchronous logic	同步时序逻辑	are derived from	源于

2.9.3 Notes

[1] Programmable logic is the means by which a large segment of engineers implement their custom logic, whether that logic is a simple I/O port or a complex state machine. 本句为主从复合句,主句为“Programmable logic is the means”,“by which”后为修饰“means”的一个定语从句,从句自身又有一个由“whether”引导的从句,修饰“implement”的状况。全句可译为“很多工程师都采用可编程逻辑来实现用户要求的逻辑功能,无论逻辑是简单的 I/O 端口还是复杂的状态机。”

[2] It can be shown through Boolean algebra that any logical expression can be represented as an arbitrarily complex sum of products. 这是一个主从复合句,“It”为形式主语,真正的主语为“that”引起的一个名词从句。全句可译为“利用布尔代数我们可以将任意的逻辑表达式表示为乘积之和的形式。”

[3] Therefore, by providing a programmable array of AND/OR gates, logic can be customized to fit a particular application. 这是一个简单句。“by providing a programmable array of AND/OR gates”为介词短语做状语,其中“providing”为动名词,其后为它的宾语,一起做“by”的介词宾语。全句可译为“因此通过提供可编程与/或阵列,可以定制逻辑来实现特定的应用。”

[4] When the connection is blown, or programmed, the AND input is disconnected from the input term and pulled high to effectively remove that input from the Boolean expression. 这是一个典型的主从复合句。“the AND input”后面为主句部分,前一部分“When the connection is blown, or programmed”为时间状语从句。全句可译为“当连接点被熔断,或称为被编程后,与门的输入不再同输入项相连并被拉到高电平,从而有效地将该输入排除在布尔表达式之外。”

[5] However, selecting flip-flops, OR/NOR polarities, and input/output configurations is performed by programming a configurable I/O and feedback structure

called a macrocell. 本句结构上只是一个简单句。“selecting flip-flops, OR/NOR polarities, and input/output configurations”为动名词短语做主语,谓语动词是“perform”。“called a macrocell”为过去分词短语做后置定语,修饰“I/O and feedback structure”。全句可译为“但是选择触发器、或/或非的极性以及输入/输出配置都是通过逻辑宏单元的编程来实现的。”

[6]The 16V8 provides up to 10 dedicated inputs that feed the AND array, whereas the 22V10 provides 12 dedicated inputs. 在这个复合句中,“whereas”引起的从句表示转折,前面主句中,注意“that feed the AND array”为“inputs”的定语从句。全句可译为“16V8 提供 10 个输入反馈给与阵列,然而 22V10 则提供 12 个输入反馈给与阵列。”

2.9.4 Reference Translation

可编程逻辑器件

很多工程师都采用可编程逻辑来实现用户要求的逻辑功能,无论逻辑是简单的 I/O 端口还是复杂的状态机。大多数可编程逻辑都是用某种类型的 HDL 来实现的,这使得工程师们从每次设计一个新的逻辑电路就要建立和化简一次布尔表达式的工作中解放出来。可编程逻辑的好处还包括以有限的工具和支持的成本快速的实现定制逻辑。

灵活的可编程逻辑产品的普适性使它为很多个人或小公司提供了定制逻辑设计的能力,他们没有充足的财力和人力资源去构建全定制的 IC。可编程逻辑器件的容量,工作电压和速度等有多种类型可供选择,几乎可以保证对于所有的应用都有十分适用的器件。选用器件时需要做些调查,因为每一个制造商在产品的特性和种类上会都有一些细微的差别。

PLD 的主要基本类型包括通用阵列逻辑器件 GAL, GAL 是在现在基本上已经过时的老可编程逻辑阵列逻辑 PAL 的结构的基础上改进而来的。虽然术语 PAL 现在仍然被广泛使用,但是在使用这个术语时它实际上所指的是 GAL 器件以及其他 PLD 器件。GAL 覆盖了 PAL 的功能,当然可以完成 PAL 所有能干的工作,所以将 PAL 淘汰。GAL 具有结构小、价格低、容易使用等优点,许多厂商,如 Cypress、Lattice 和 Texas Instruments 等,都在生产 GAL 产品。利用布尔代数我们可以将任意的逻辑表达式表示为乘积之和的形式。因此通过提供可编程与/或阵列,可以定制逻辑来实现特定的应用。GAL 器件有一个很大的可编程与门阵列,接入所有输入项,如图 2.9.1 所示。

每个输入变量及其反变量都可连接到与门上。与门阵列的输出(积)连接到独立的或门(和),从而形成用户定义的布尔表达式。

每根与门的垂直线和水平输入线的交接处都是可编程的。在 PLD 发展的早期,连接点都是由熔丝组成的,通过相当于高电压熔断熔丝的办法来生成所要的器件。熔丝器件是不可重复编程的,一旦熔丝熔断后,它就不可恢复。如今主要是通过 EEPROM 技术和 CMOS 开关器件来保证非易失性,从而实现非挥发性再编程。但是由于历史原因,熔丝这个名称仍然被使用。

连接点处的默认配置是有一个完好的熔丝，所以输入项是连接到与门输入上。当连接点被熔断，或称为被编程后，与门的输入不再同输入项相连并被拉到高电平，从而有效地将该输入排除在布尔表达式之外。定制一个 GAL 可编程与门的原理可用图 2.9.2 来说明。

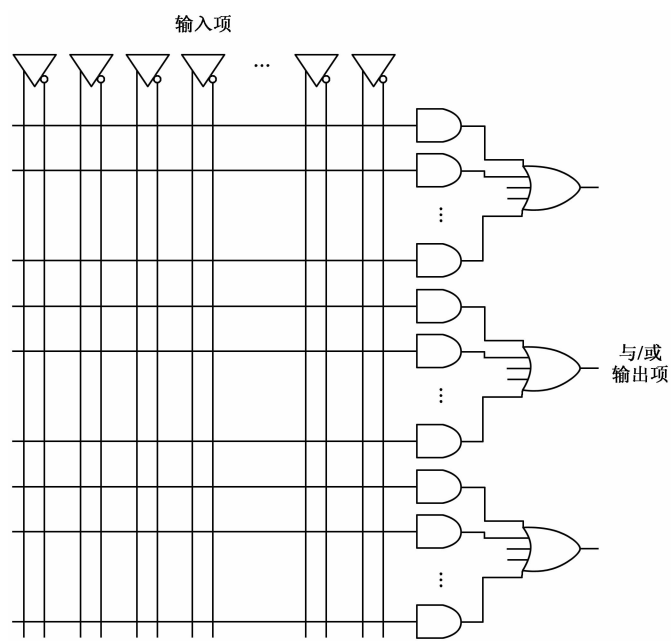


图 2.9.1 GAL 与/或结构

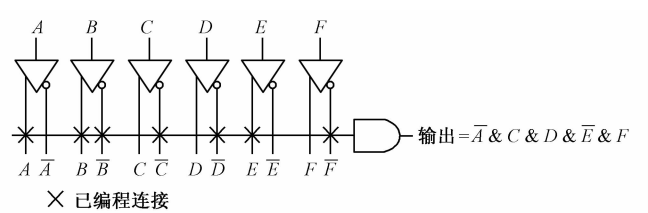


图 2.9.2 编程与输入项

与阵列具有充分的可编程性,或连接的连线却是固定的。每个 GAL 器件将不同数目的与项接到或门上。如果几个与项在某个特定的布尔表达式中并不需要,可以通过这将这些与门的输出置 0 来有效地将其禁止。这是通过不对这些与门输入进行编程来实现的。必须注意连接到与门的输入包括原变量和反变量。当所有的与门连接点都未熔断,结果积表达式 $A \& \bar{A} = 0$,因此将该门输出结果置 0,从而保证该项不参与逻辑或运算。

GAL 逻辑定制的主要工作就是对阵列进行编程。但是选择触发器、或/或非的极性以及输入/输出配置都是通过对称为宏单元的编程来实现的。宏单元的基本概念就是最

终决定如何处理与/或表达式,以及与宏单元相关的引脚是如何运作的。

不同 GAL 中的宏单元可能有些细微的不同,图 2.9.3 所示的是一个 GAL 宏单元的基本原理的图解。无论某项是被寄存,还是反馈信号直接来自触发器输出或来自该引脚,最终或/或非项的结果都由多路选择器来决定。通过配置宏单元的输出可决定引脚的行为。

虽然也有其他类型的器件,现在通用的 GAL 器件有两种:16V8

和 22V10。它们分别包含 8 个和 10 个宏单元。16V8 提供 10 个输入给与阵列,然而 22V10 则提供 12 个输入反馈给与阵列。22V10 中的有一个输入也用作所有的触发器的时钟控制信号,可以在宏单元中选用。22V10 中的每个宏单元都可以通过一个专用的与项来独立控制输出使能。16V8 的灵活性则相对差一些,因为它不能根据器件的配置情况,任意地反馈宏单元的输出。而且当配置成寄存器模式时,宏单元中的触发器有效,这时得占用两个专用的引脚用于时钟和输出使能控制。

按照现在的标准,GAL 器件的密度相当低。由于规模较小,因此它具有低造价和高速度等优点。在 GAL 上实现特定的逻辑功能包括以下几个步骤。首先,用图形原理图或文字(HDL)来表示逻辑,然后利用转换工具或综合工具将它转化为一个网表形式,最后在可编程与门阵列上实现具体函数功能,从而在目标器件上实现网表。对于给定 GAL 上固定的与/或阵列,利用适配软件可以将任意的布尔表达式转换成乘积-和的形式。处理的结果是形成一个程序映像,也称为熔丝映像图,它严格确定了哪些连接点或熔丝要编程,哪些保持默认态。程序映像也可包括宏单元的配置以及设备相关的特性等其他信息。

现在的 PLD 开发软件在大多数情况下可以自动进行后端 GAL 综合处理,不需要人为的干预。简单直接的与门阵列可编程逻辑减少了为实现布尔表达式而要做的繁杂的逻辑组配过程,使其变得直观明了。输入信号通过引脚和端口传输给与阵列,仅仅通过两个逻辑门,它或者输入到宏单元的触发器,或者直接通过一个 I/O 引脚输出。GAL 中的逻辑块都是紧密地连接在一起的,所以器件尺寸很小,从而保证了内部较小的传输延迟。由于信号以较小传输延迟在明确的路径上传输,这些特性使得 GAL 结构具有非常快速的时间特性。

GAL 是一种具有很强的前期预测能力的逻辑实现技术。如果一个 GAL 无法实现逻辑需求,那么我们或者通过优化算法,或者通过将逻辑任务分解,利用多个器件,共同完成设计任务。如果 GAL 能够满足逻辑功能,但不能满足时序要求,我们或者通过优化算法,或者采用高速器件来实现设计要求。由于 GAL 逻辑组配及结构的简单性,不能要求它做像复杂的 CPLD 做的同样的工作。但这不能认为是 GAL 缺少灵活性。GAL 能够

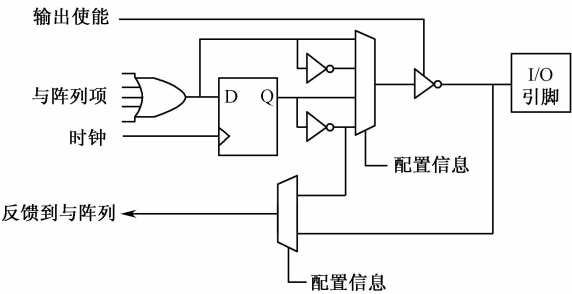


图 2.9.3 GAL 宏单元

完成它该完成的功能,倒是工程师必须合理地运用这项应用技术来解决手头上的问题。GAL 结构简单是它的最大优势。

Lattice Semiconductor 的 GAL22LV10D-4 在最糟的情况下的最大输入输出延迟时为 4 ns,这一延时指标使该器件可以用于快速微处理器接口的地址译码。同样的 22V10 器件 $t_{co} = 3$ ns,频率最大为 250 MHz。参数 t_{co} 规定的是引脚到引脚的特性,其中包括时钟延迟、通过输入引脚的输入信号的延迟以及通过输出引脚的输出信号的延迟。内部触发器实际具有较快的 t_{co} ,这对内部逻辑反馈通路是很重要的。最高的时钟频率是所有 PLD 器件和其他一些考虑的感兴趣的方面。它们定义了触发器的最高计时频率,但是同步时序分析表明讨论 f_{MAX} 比较重要,尤其要重视逻辑的传输延迟和触发器间的相互连接。GAL 结构中确定和快速的反馈通路减少了内部的传输延迟。GAL22LV10D 的反馈延迟是 2.5 ns,包括实际触发器的 t_{co} 以及通过与/或阵列反馈的信号传输延迟。反馈延迟加上触发器的 3 ns 的 t_{SU} , f_{MAX} 为 182 MHz,可用于大多数通用的包含反馈通路的同步逻辑设计中,如状态机等。

2.9.5 Reading Materials

Hardware Description Languages

Many basic peripheral logic functions are available in off-the-shelf ICs. A variety of UART ICs are available, DMA controllers are available, and simple address decoding can be accomplished with 7400 devices such as the 74LS 138. As digital systems grow more complex, the chances increase that suitable off-the-shelf logic will be either unavailable or impractical to use. The answer is to design and implement custom logic rather than relying solely on a third party to deliver a solution that does exactly what is needed.

Logic design techniques differ according to the scale of logic being implemented. If only a few gates are needed to implement a custom address decoder or timer, the most practical solution may be to write down truth tables, extract Boolean equations with Karnaugh maps, select appropriate 7400 devices, and draw a schematic diagram. This used to be the predominant means of designing logic for many applications, especially where the cost and time of building a custom IC was prohibitive. The original Apple and IBM desktop computers were designed this way, as witnessed by their rows of 7400 ICs.

When functions grow more complex, it becomes awkward and often simply impossible to implement the necessary logic using discrete 7400 devices. Reasons vary from simple density constraints—how much physical area would be consumed by dozens of 7400 ICs—to propagation delay constraints—how fast a signal can pass through multiple discrete logic gates. The answer to many of these problems is custom and semicustom logic ICs. The exact implementation technology differs according to the

cost, speed, and time constraints of the application, but the underlying concept is to pack many arbitrary logic functions and flip-flops into one or more large ICs. An application specific integrated circuit (ASIC) is a chip that is designed with logic specific to a particular task and manufactured in a fixed configuration. A programmable logic device (PLD) is a chip that is manufactured with a programmable configuration, enabling it to serve in many arbitrary applications.

Part 3 Microprocessors and Distributed Computer Control

3.1 A Brief History of the Microprocessor(I)

3.1.1 Text

Abstract: The following article describes the evolution of the microprocessor which focused on the technology, the companies and the people behind the invention. It begins with a brief history before the Intel 4004, then describes the designing of the chip. It follows the microprocessor through its iterations to RISC, parallel processing and to today's super-RISC designs.

Background

The first Integrated Circuit(IC) was invented in late 1958 by Jack S. Kilby working for Texas Instruments. The company was an innovative manufacturer of transistors and Kilby's first job with the company was designing micro modules for the military. This involved connecting many germanium wafers of discrete components together by stacking each wafer on top of one another.

Connections were made by running wires up the sides of the wafers. Kilby saw this process as unnecessarily complicated. He realized that if a piece of germanium was engineered properly, it could act as many components simultaneously. Thus the first IC was born. A year later, Fairchild Semiconductor (founded in 1957), a division of Fairchild Camera & Instrument Corporation invented the modern silicon diffusing process, or planar process which is still used today. The IC process gradually evolved over the next ten years including moving the development process over to computer aided design in 1967^[1].

In June 1968, Robert Noyce (who had helped in developing Fairchild's silicon process), Gordon Moore and Andrew Grove resigned from Fairchild and founded their own company. Intel (short for Integrated Electronics) was born. The departure of the three men was significant not least because Robert Noyce was a co-founder and vice president of Fairchild.

The reason behind the departure was the skepticism of the Fairchild managers towards the future of the integrated circuit. Thus Fairchild's subsidiary semiconductor operation resented the managers as they felt the invention had a great deal to offer.

The Reasons Behind Producing an Integrated CPU

Busicom, a trading name of a now defunct Japanese company called ETI, was

planning a range of next generation programmable calculators. Busicom had designed 12 chips and asked Intel to produce them. This was as a result of Intel's expertise with high transistor count chips. Marcian Hoff Jr. was assigned to the project and after studying the designs concluded its complexity far exceeded the usefulness of a calculator^[2]. Hoff was able to contrast the design with the DEC Program Data Processor 8 (PDP-8) ^[3]. The PDP-8 had a relatively simple architecture, yet could perform very high level operations. Hoff realized a general purpose processor would be a simpler design, yet able to handle a greater number of tasks. The MCS-4 chipset and in particular, the 4004 integrated CPU were thus conceived. In 1969 Busicom chose Intel's "Microcomputer on a chip" (the word microprocessor wasn't used until 1972) over its own 12 chip design. Busicom's contract with Intel stated Busicom had exclusive rights to buy the new chip set (CPU, ROM, RAM, I/O), however Intel agreed with Busicom in 1971 that in exchange for cheaper chip prices, Intel would have full marketing rights enabling them to sell the chips to whoever wanted to buy^[4].

Intel CPU Design to the 8086

In late 1969, after the 4004 instruction set had been defined, Computer Terminals Corporation (CTC) asked Intel to develop an LSI registers chip for a new intelligent terminal they were developing. Due to experience with the 4004 and the furious pace of development within the industry, Stan Mazor (who had helped on the 4004) and Hoff agreed that they would put the complete processor on one chip. The 8008, an 8 bit processor was defined and work began immediately. The chip was rejected by CTC as it required many support chips (a minimum of 20 TTL packages for memory and I/O) and was too slow. Chip design continued in parallel to the MCS-4 and in April 1972 the 45 instruction CPU was launched.

The chip became a great success. Intel looked at the CTC rejection of the 8008 and realized they had to make a general purpose processor requiring only a handful of support chips. The Intel 8080 was born in April 1974 even though it was announced earlier. Intel did this to give customers sufficient lead times to design the chip into their products. The 8080 had 4,500 transistors, twice the number in the 4004 and could address 64K bytes of memory. Its speed was mainly down to the use of electron doped technology as opposed to hole doped MOS transistors. The chip was an astounding success and became the industry standard, emulated by other companies.

In 1978, Intel produced its first 16 bit processor, the 8086. It was source compatible with the 8080 and 8085 (an 8080 derivative). This chip has probably had more effect on the present day computer market than any other, although whether this is justified is debatable^[5]; the chip was compatible with the 4 year old 8080 and this

meant it had to use a most unusual overlapping segment register process to access a full 1 Megabyte of memory.

The Early Years: Not Just Intel

Although Intel had invented the microprocessor and had grown from a three man start-up in 1968 to an industrial giant by 1981 with 20,000 employees and revenues of 188 million, they were not the only company developing microprocessors. By July 1974, 19 microprocessors were either available or announced. By 1975, that number increased to 40 and by 1976 it was 54. Late 1972 saw the second ever processor, with Rockwells PPS-4, a 4 bit processor. Another 4 bit processor, the Texas Instruments TMS 1000 was introduced on the market in 1974, although it had been designed in 1971. This was around the same time as Intel's 4004, but TI failed to realize the potential, and left the TMS 1000 to spend its first three years controlling a Texas Instruments calculator. Surprisingly, the TMS 1000 was also the first microcontroller, as it contained its own RAM and ROM on chip.

By the late 1970s, the cost of the chip had fallen to a few dollars, and had become the processor of choice for consumer electronics. It was being produced in over forty variants and sold in the hundreds of millions. The staggering development in the field was also exemplified in 1974 by the National Semiconductor Processing And Control Element (PACE). National was a Fairchild offshoot and thus had a large skills base. Unfortunately, the chip was designed using hole doped MOS transistors. This resulted in a third of the speed of the chip if instead it had been designed using electron doping^[6].

Clones

Due to the success of the Intel 8008, Zilog and Motorola produced competing chips. Motorola realized the potential of the microprocessor after seeing the 8008. In mid 1974 they launched the 6800, a processor in the same market as the 8080. Motorola launched the 6800 with a wide variety of support in the way of system oriented hardware. This integration proved a major factor in the popularity of the 6800, as it did not have Intel compatibility to fall back on^[7]. Popular as the chip was, it fell well short of a derivative designed by a group of engineers who left Motorola to begin their own company^[8]. MOS Technologies delivered their 6800— influenced processor, the 6502 in 1975.

The 6502 was successful due to simple design, single power line and cheapness. It became a favorite for the emerging small home computer businesses including Apple, Commodore and Acorn. Being a simple yet powerful design it was able to hold its own against the later designed and more powerful Z80^[9]. As a result, it had an influence on the concept of Reduced Instruction Set Computing (RISC) and especially the Acorn ARM processor.

3.1.2 Specialized English Words

chip	芯片	chip set	芯片组, 一组芯片
iteration	重复(物), 反复, 迭代(法), 逐步逼近法	furious	高速的, 猛烈的, 狂暴的
RISC	精简指令集计算机 (Reduced Instruction Set Computer 之缩写)	electron doped technology	电子掺杂 技术
parallel processing	并行处理	hole doped	空穴掺杂的
Integrate Circuit	集成电路	astounding	令人震惊的
transistors	晶体管	source compatible	源(代码)兼容的
micro models	微型化模块	start up	创业阶段, 启动时期
germanium	锗	revenues	总收入, 财源
wafers	晶(圆)片, 薄片, 薄脆饼	microcontroller	微控制器
discrete components	分立元件	RAM	随机访问存储器 (Random Access Memory 之缩写)
stacking	堆叠	ROM	只读存储器 (Read Only Memory 之缩写)
engineer	工程师, 精心策划, 设计制造	on chip	在片的, 片内的
silicon diffusing process	硅扩散工艺(法)	staggering	难以想象的, 骇人的
planar process	平面工艺(法)	be exemplified	成为典范, 例证
computer aided design	计算机辅助设计, CAD	clones	克隆, 复制品, 仿制品, 类似品
skepticism	怀疑态度, 怀疑主义	fall back on	求助于
subsidiary	辅助的, 次要的	derivative	衍生物, 派生物
resent	愤恨	emerging	正在出现的, 初露头角的
have...to offer	能提供, 满足要求	single power line	单一供电渠道, 单一电源
defunct	已不再使用的	hold ... against	以……和……对抗, 因……而贬低(某人)
expertise	专长		
conceive	构想, 设想, 怀孕		

3.1.3 Notes

[1] The IC process gradually evolved over the next ten years including moving the development process over to computer aided design in 1967. 这是一个主谓结构的简单句式, 主语为“IC process”, 谓语动词为“evolved”。其余为“evolved”的三个状语, 一是副词“gradually”, 二是介词短语“over the next ten years”, 三是“including...”。注意第三个成分中“moving...”是动名词, 和后续部分一起做“including”的宾语。全句可译为“IC 工艺在接下来的十年中不断发展, 包括 1967 年发展出来的计算机辅助设计技术。”

[2] Marcian Hoff Jr. was assigned to the project and after studying the designs

concluded its complexity far exceeded the usefulness of a calculator. 这是一个并列复合句,“was assigned”和“concluded”共一主语“Marcian Hoff Jr...”,而“concluded”的宾语为略去引导词“that”的宾语从句。全句可译为“小 McCain Hoff 被安排开发这个项目。经过研究,Hoff 得出结论,发现它将远远超过一个计算器所需要的复杂程度。”

[3]Hoff was able to contrast the design with the DEC Program Data Processor 8 (PDP-8). “contrast ...with”意为两者相比。全句可译为“Hoff 把它和 DEC 公司的 PDP-8 处理器进行了比较,发现两者差距甚远。”

[4]Busicom's contract with Intel stated Busicom had exclusive rights to buy the new chip set (CPU, ROM, RAM, I/O), however Intel agreed with Busicom in 1971 that in exchange for cheaper chip prices, Intel would have full marketing rights enabling them to sell the chips to whoever wanted to buy. 这是一个成分较复杂的复合句。“however”前后是两个并列的主句,前一句的谓语动词“stated”后跟略去了“that”的宾语从句“Busicom had exclusive rights to buy the new chip set (CPU, ROM, RAM, I/O)”;后一句的谓语动词是“agreed”,后跟宾语从句“that in exchange...”,从句中“whoever wanted to buy”=“any person who wanted to by”,是一个名词从句,做“to”的介词宾语。全句可译为“Busicom 与 Intel 所签的合同言明 Busicom 拥有这个新芯片 (CPU,ROM,RAM,I/O) 的独家购买权,然而到 1971 年,为了降低芯片价格,Intel 与 Busicom 达成一致,Intel 有向任何有意向的买家出售 4004 芯片的充分销售权。”

[5]This chip has probably had more effect on the present day computer market than any other, although whether this is justified is debatable. 句中“whether this is justified”是从句“although...”的主语从句,注意从句和从句中的主语从句都是系动词加表语的结构,所以出现“is justified”和“is debatable”连到一起的情况。全句可译为“下面的说法是否公正也许还值得争论,但这款芯片对今天计算机市场的影响大概比任何其他芯片都大。”

[6]This resulted in a third of the speed of the chip if instead it had been designed using electron doping. 句中“if”引起虚拟条件从句。“instead”做状语,为现在分词短语做状语。“resulted in”意为“结果是”。全句可译为“这使其速度只有假如采用电子掺杂的设计制造出的芯片的速度的三分之一。”

[7]This integration proved a major factor in the popularity of the 6800, as it did not have Intel compatibility to fall back on. 句中“as”引起的状语从句,表示原因和条件,“integration”在句中指上句的“a wide variety of support”。“prove”一词指“事实或时间终于证明”。全句可译为“事实证明,这种广泛的支持是 6800 在没有与 Intel 的兼容性可资利用的情况下得以流行的主要原因。”

[8]Popular as the chip was, it fell well short of a derivative designed by a group of engineers who left Motorola to begin their own company. 句中“Popular...was”是状语从句“as the chip was popular”的倒装形式,“as”在这里等于“although”,意为“尽管”。“well short of”意为“严重缺乏”。“designed by...”为“derivative”的后置定语。“who...”为

“engineers”的定语从句。全句可译为“尽管这款芯片很流行,但它大大缺乏后续产品。这款芯片的设计师离开了 Motorola,创建了自己的公司。”

[9]Being a simple yet powerful design it was able to hold its own against the later designed and more powerful Z80. 句中“Being...design”为现在分词短语,用做状语。全句可译为“由于它的设计简洁明快,功能强大,足以和后来设计的更强大的 Z80 相抗衡。”

3.1.4 Reference Translation

微处理器简史(I)

摘要:本文讲述微处理器的发展历程,着重相关的技术、公司及人员。文章先简单介绍了 Intel 4004 之前的情况,然后讲述 4000 芯片的设计。接下来介绍各种后续的微处理器,直到 RISC、并行处理及今天的超级 RISC 设计产品。

背景

第一块集成电路(IC)是由在德州仪器公司工作的 Jack S. Kilby 于 1958 年年末发明的。德州仪器公司当时是一家创新型的晶体管制造商,而 Kilby 在公司的第一项工作是为军方设计微型化模块器件。这涉及把分立元件构成的很多锗晶片一片片堆叠到一起。必须在晶片周边向上进行布线连接。Kilby 认为这样复杂的过程是完全没有必要的。他意识到只要精心设计布局,一块锗晶片就能同时起到很多元件所起的同样的作用。就这样,第一块集成电路(IC)诞生了。一年后,仙童半导体公司(成立于 1957 年)——仙童照相机与仪器公司的一个部门——发明了现代硅扩散工艺,即迄今仍在使用的平面工艺。IC 工艺在接下来的十年中不断发展,包括 1967 年发展出来的计算机辅助设计技术。

1968 年 6 月,Robert Noyce(他协助开发了仙童公司硅扩散工艺)、Gordon Moore 和 Andrew Grove 从仙童公司辞职,组建了自己的公司。Intel(Integrated Electronics 之缩写,集成电子电路公司)成立了。这三人离开仙童的意义非同小可,尤其因为 Noyce 是仙童的创始人之一及副总裁。

三人离走的背后原因是由于仙童公司管理层对集成电路的未来所持的怀疑态度。他们认为这项发明花费会太大,结果引起下属的这个半导体子公司感到愤愤不平。

集成制造 CPU 的内在原因

Busicom 是一家已不存在的日本公司 ETI 的一个商业项目的名称,打算推出一系列新一代的可编程计算器。它设计了 12 种芯片,要求 Intel 进行生产。这可是 Intel 所擅长的大数量晶体管芯片制造技术。

小 McCain Hoff 被安排开发这个项目。经过研究,Hoff 得出结论,发现它将远远超过一个计算器所需要的复杂程度。Hoff 把它和 DEC 公司的 PDP-8 处理器进行了比较,发现两者差距甚远。PDD-8 的结构相对简单,而运行水平达到很高的高度。Hoff 认识到一个通用的处理器的设计可以并不那么复杂但却能承担更繁重的任务。

MCS-4 芯片组,特别是 4004 集成 CPU 就这样构思出炉了。1969 年,Busicom 从 Intel 的 12 个设计中选中了“芯片上的微型计算机”(微处理器一词到 1972 年才使用)这一款设计。Busicom 与 Intel 所签的合同言明 Busicom 拥有这个新芯片(CPU、ROM、RAM、I/O)的独家购买权,然而到 1971 年,为了降低芯片价格,Intel 与 Busicom 达成一致,Intel 有向任何有意向的买家出售 4004 芯片的充分销售权。

Intel 设计出 8086

临近 1969 年年末,4004 的指令集业已确定。计算机终端公司(CTC)要求 Intel 为他们正在设计的一种新智能终端设备开发一款大规模集成电路寄存器芯片。由于有了 4004 的经验,加上这个领域迅猛的发展,Stan Mazor(他参与了 4004 项目)和 Hoff 承诺将整个处理器制作在一块芯片上。一个 8 位的处理器 8008 就这样确定下来并立即着手开始工作。但这款芯片遭到 CTC 的拒绝,原因是需要太多的支持芯片(至少需要为存储器和 I/O 配 20 个 TTL 芯片),而且速度太慢。Mazor 和 Hoff 继续设计 8008 芯片,MCS-4 也同时在开发之中。到了 1972 年 4 月,这个拥有 45 条指令的 CPU 发布了。

这个芯片是一个巨大的成功。面对 CTC 对 8008 的拒绝,Intel 认识到应当开发一种只需要很少外用支持芯片的通用处理器。Intel 8080 于 1974 年 4 月诞生,尽管它的发布要早一些。Intel 这样做,给了用户充分的设计时间将 8080 纳入各自的产品体系。8080 有 4500 个晶体管,是 4004 的两倍,并能访问 64 KB 的存储器。它的工作速度主要由于采用了电子掺杂工艺而非空穴掺杂工艺的 MOS 晶体管而较慢。这款芯片是一项令人震撼的成功,并成为工业标准,为其他公司所仿效。

1978 年 Intel 制成第一片 16 位 CPU——8086。它与 8080 及 8085(8080 的衍生品种)源代码兼容。下面的说法是否公正也许还值得争论,但这款芯片对今天计算机市场的影响大概比任何其他的芯片都大。8086 和四年前的老芯片兼容,这意味着 8086 得采用一种非常独特的段寄存器重叠处理方法来访问全部 1 MB 存储器空间。

早期年代:不仅是 Intel

尽管 Intel 发明了微处理器,而且从 1968 年草创时期的三个人成长为到 1981 年拥有 20 000 名雇员、总收入达到 1.88 亿美元的工业巨头,但它并非开发微处理器的唯一企业。截至 1974 年 7 月,共有 19 种微处理器投入使用或者得以发布。到 1975 年,这个数字增加到 40,而到 1976 年成了 54。1972 年晚些时候,出现了第二个微处理器,即 Rockwell 的 PPS-4,这是一种 4 位的处理器。另外一个 4 位的处理器,Texas Instrument 的 TMS 1000,设计于 1971 年,但到 1974 年才推向市场。这和 Intel 的 4004 差不多是同时代的,但 TI 没有认识到 TMS 1000 的潜在价值,而是让它在头三年里仅仅用于 TI 出品的一款计算机。令人惊奇的是,TMS 1000 还是第一个微控制器,具有片内 RAM 和 ROM。

到了 1970 年代后期,微处理器的成本降至一片几美元,成为消费电子产品的首选,品种超过 40 种,销售量上亿。这一领域的惊人发展到 1974 年由国家半导体发展与控制集团(PACE)再次重复上演。PACE 是仙童的分支部门,因而拥有强大的技术基础。遗憾

的是,它的产品采用了空穴掺杂 MOS 管,这使其速度只有假如采用电子掺杂技术设计制造出的芯片的速度的三分之一。

仿效品

由于 Intel 8008 的成功,Zilog 和 Motorola 也推出了各自的竞争性产品。Motorola 在见到 8008 之后认识到了微处理器的潜在价值。1974 年年中,Motorola 推出了 6800,面向和 8080 一样的市场。6800 拥有面向系统的各种各样硬件的支持。事实证明,这种广泛的支持是 6800 在没有与 Intel 的兼容性可资利用的情况下得以流行的主要原因。尽管这款芯片很流行,但它大大缺乏后续产品。这款芯片的设计师离开了 Motorola,创建了自己的公司。

1975 年 MOS Technologies 公司发布了他们的颇受 6800 影响的 6520 芯片。

6502 的成功在于设计简洁、单电源供电以及价格低廉,受到包括 Apple、Commodore 和 Acorn 在内的正在初露头角的小型家用电脑制造商的青睐。由于它的设计简洁明快,功能强大,因而足以和后来设计的更强大的 Z80 相抗衡。作为一项成果,6502 对精简指令集计算机(RISC)的理念产生了影响,尤其是影响了 Acorn 的 ARM 处理器。

3.1.5 Reading Materials

A Timeline of Microprocessors (I)

This document is intended to provide an idea of who preceded (先于) [and superseded(取代)]who. Please message(发送信息给……) me comments, corrections, and so on.

The Seventies

The seventies were a good time for intel, mostly because they were the first players in the game. Motorola jumped in rapidly thereafter, however, and brought out the ubiquitous(无处不在的) 6800 and later the even more important 68000 during the same time frame(时间段). Even today, however, 808x CPUs are more popular in embedded systems than the more powerful Motorola 68000, if for no other reason than inertia(惯性, 习惯). Intel got there first, and got the ball rolling(拔得头筹并保持发展势头). A great deal of their installed base comes from the fact that the IBM PC and every clone(克隆, 仿效品) of it thereafter carried an intel CPU.

IBM also invented the first RISC CPU during this decade (barely). They began work on it way back in 1975. That chip was never released but concepts in its design made it all the way into the PowerPC by way of ROMP and then POWER.

1971: 4004 (Intel) Used in the Busicom calculator. First microprocessor. 4 bits, 2300 transistors, 740 kHz, 0.06 MIPS.

1972: 8008 (Intel) Used in the Mark-8.

1974: 8080 (Intel) Used in the Altair.

1974: MC6800 (Motorola) Easier to implement than Intel 8080 as it needs only one voltage and no support chips to operate. Mostly sold for peripheral and industrial control.

1975: MC6501 (MOS) Pin-compatible with Motorola MC6800, leading to a lawsuit against MOS.

1975: MC6502 (MOS) Replaces the 6501, and is not pin-compatible with MC6800. Used in Apple 2 and Commodore VIC 20. MOS Technology was purchased by Commodore later the same year.

1976: 8085 (Intel) Improved version of the 8080; uses only +5 V, where 8080 needs several voltages, and with additional instructions as well.

1976: TMS9900 (TI) First 16 bit microprocessor.

1976: Z80 (Zilog) The Z80's instruction set is a superset(超级指令集) of the Intel 8080. It later becomes one of the most ubiquitous embedded processors of all time. The de facto(实际存在的) standard for computers running CP/M. Also featured in the Radio Shack TRS-80 and the Nintendo Game Boy, among many others.

1978: 8086 (Intel) Used (later) in the IBM PC. Also, the complementary 8087 math coprocessor.

1979: 8088 (Intel) Cost reduced 8086, with an 8 bit bus instead of 16 bit.

1979: 801 (IBM) First RISC CPU made. Never commercialized(从未商品化, 从未上市).

1979: Z8000 (Zilog) 16 bit chip.

1979: MC68000 (Motorola) 16 bit processor with 24 bit addressing.

3.2 A Brief History of the Microprocessor (II)

3.2.1 Text

The Zilog chip, the Z80 was significant in that it was compatible with the 8080 yet added 80 more instructions^[1]. However this compatibility was not unexpected as Zilog was founded by engineers who had left Intel. Two of those engineers were Frederico Faggin and Masatoshi Shima who had designed the 4004 and 8080 for Intel. The Zilog (an acronym in which the Z stands for “the last word,” the “i” for integrated and “log” for logic) Z80 was a very powerful processor including on-chip dynamic memory refresher circuits. This enabled system designers such as Sir Clive Sinclair to produce computers with very little extra circuitry and hence at very little cost.

A year after Intel produced their first 16 bit processor, Motorola introduced another influential and long lived chip, the 68000. It was able to address a massive 16 megabytes and was able, through intricate internal circuitry to act like a 32 bit processor internally. The chip found fame in the Macintosh, Amiga and Atari personal computers.

A new philosophy—RISC

Most commentators see RISC as a modern concept, more akin to the 1990s, yet it can be traced to 1965 and Seymour Cray's CDC (Control Data Corporation) 6600. RISC design emphasizes simplicity of processor instruction set, enabling sophisticated architectural techniques to be employed to increase the speed of those instructions. A classic example is the VAX architecture where the INDEX instruction was 45% to 60% faster when replaced by simpler VAX instructions^[2]. The CDC 6600 has many RISC features including a small instruction set of only 64 op codes, a load/store architecture and register to register operations. Also, instructions weren't variable lengths, but 15 or 30 bits long.

Although the term RISC was not used, IBM formalized these principles in the IBM 801(1975), an Emitter Coupled Logic (ECL) multichip processor. The architecture featured a small instruction set, load/store memory operations only, 24 registers and pipelining.

When RISC became popular in the late eighties, IBM tried to market the design as the Research OPD (Office Products Division) Mini Processor (ROMP) CPU, but it wasn't successful. The chip eventually became the heart of the I/O processor for the IBM 3090. The term RISC first came from one of two University research projects in the USA. The Berkeley RISC I formed the basis for the commercial Scalable (formerly Sun) Processor Architecture (SPARC) processor, whilst Stanford University's Microprocessor without Interlocked Pipeline Stages (MIPS) processor was commercialized.

Around the time the results of the Stanford and Berkeley projects were released, a small UK home computer firm, Acorn was looking for a processor to replace the 6502 used in its present line of computers. Their review of commercial microprocessors including the popular 8086 and 68000 concluded that they were not advanced enough, so in 1983 began their own project to design a RISC microprocessor. The result, ARM (Advanced RISC Machine, formerly Acorn RISC Machine) is probably the closest to true RISC of any processor available.

Parallelism—The Transputer

In 1979, Inmos was formed by the British government to produce innovative silicon based products competing on the world stage. The formation was partly in response to the increasing dominance of the market by the USA and the need to provide the UK with manufacturing facilities. During the summer of 1980, Inmos were working on its first microprocessor, however events were not smooth with two engineers having inflexible positions over their idea of the architecture for this microprocessor^[3]. David May who

had been recruited from Warwick University and Robert Milne who had come from Seicon, a specialist company producing complex computer programs were the engineers. Milne felt that the Transputer, the name given for the Inmos chip, should be the first chip in the world specially tailored to run Ada. He felt this was the future of micro-processor design which was in strict contrast to May and Tony Hoare. Hoare was an academic guru from Warwick where he had worked with May and shared a simplistic approach to the Transputer design. Iann Barron, who had been the driving force behind Inmos became tired of this rambling and forced his view on the team. His views happened to encompass those of May but he also envisaged the multiplicity of individual processors all working concurrently. The transputer came to market in 1985 as the T-212, a 16 bit initial version with a RISC like instruction set. Each chip uniquely had 4 serial links which enabled the microprocessor to be connected to other Transputers in a network. In 1994, the T-9000 was launched. It is a design optimized for use in parallel computers using a systolic array configuration.

The SuperRISCs

In 1988, DEC formed a small team that would develop a new architecture for the company. Eleven years previously, it had moved from the PDP-11 to the VAX architecture, but it was seen that it would start lagging behind by the early 1990s^[4]. The project turned out to be huge with more than 30 engineering groups in 10 countries working on the Alpha AXP architecture as it came to be known.

The team were given a fabulous design opportunity; an architecture that would take DEC into the 21st century. To accommodate the 15-25 year life span of the processor, they looked back over the previous 25 years and concluded a 1000 fold increase in computing power occurred^[5]. They envisaged the same for the next 25 years, and so they concluded that their designs would, in the future, be run at 10 times the clock speed, 10 times the instruction issue rate, (10 times superscalar) and 10 processors working together in a system. To enable the processor to run multiple operating systems efficiently, they took a novel approach and placed interrupts, exceptions, memory management and error handling into code called PALcode (Privileged Architecture Library) which had access to the CPU hardware in a way which microcode normally has. This enables the Alpha to be unbiased toward a particular computing style.

3.2.2 Specialized English Words

acronym	首字母略语	refresher circuits	刷新电路
dynamic memory	动态存储器	megabytes	兆字节, 百万字节

intricate 错综复杂的	guru 大师,权威,专家
akin to 相似	rambling 漫步,散漫离题
formalize 使……成为正式,使……定形	encompass 包括,包含,困境
emitter coupled logic(ECL) 发射极耦合逻辑	envisage 设想,展望
pipelining 流水线(工作方式)	multiplicity 多样性
whilst = while	systolic (心脏)收缩的
Microprocessor without Interlocked Pipeline Stages (MIPS) 无互锁流水级微处理器	systolic array 脉冲阵列
recruit 吸收(新成员),征募(新兵)	lag behind 滞后于,落后于
Ada 一种计算机语言	fabulous 绝妙的,极好的
	life span 生命期,使用寿命
	fold 倍乘
	issue rate 吞吐率,进出率

3.2.3 Notes

[1] The Zilog chip, the Z80 was significant in that it was compatible with the 8080 yet added 80 more instructions. 在这个复合句中,“it was…8080”是介词“in”的宾语“that”的定语从句。“that”指的“那件事情”由它的后置定语来具体表示。全句可译为“Zilog 的芯片 Z80 在保持与 8080 兼容这一点上,是具有重要意义的,它还另外新增了 80 条指令。”

[2] A classic example is the VAX architecture where the INDEX instruction was 45% to 60% faster when replaced by simpler VAX instructions. 在这个复合句中主句为系表结构。“where…”是表语“the VAX architecture”的定语从句。“when…”是从句的状语。“when replaced”系“when it was replaced”之缩写,可以视为状语从句。全句可译为“VAX 的结构就是一个典型的例子。采用了较简单的 VAX 指令后,INDEX 指令的执行速度可以提高 40%到 60%。”

[3] During the summer of 1980, Inmos were working on its first microprocessor, however events were not smooth with two engineers having inflexible positions over their idea of the architecture for this microprocessor. 句中“having…over…”为现在分词短语,做“engineers”的后置定语,意为“对……有……的(看法等)”。全句可译为“到了 1980 年夏季,Inmos 着手开发自己的第一个芯片。但是事情由于两名工程师对于微处理器的结构的设想各执一词而进展并不顺利。”

[4] Eleven years previously, it had moved from the PDP-11 to the VAX architecture, but it was seen that it would start lagging behind by the early 1990s. 句中“it was seen”中的“it”为形式主语,真正的主语是“that”引导的名词从句。全句可译为“虽然早在 11 年前,DEC 已从 PDP-11 结构转向了 VAX 结构,但是这时已经可以看出,到 1990 年代早期就会开始落后于人。”

[5] To accommodate the 15-25 year life span of the processor, they looked back

over the previous 25 years and concluded a 1000 fold increase in computing power occurred. 注意句中“concluded”后面是省去了先行词“that”的宾语从句。从句的主语是“a 1000 fold increase”,谓语动词是“occurred”。全句可译为“为了让这个微处理器能有 15 到 25 年的生命期,设计者回顾了过去 25 年的情况,结论是 25 年间计算机能力增加了 1000 倍。”

3.2.4 Reference Translation

微处理器简史(II)

Zilog 的芯片 Z80 在保持与 8080 兼容这一点上,是具有重大意义的,它还另外新增了 80 条指令。这个兼容性并不令人感到意外,因为 Zilog 是由几位离开 Intel 的工程师所创建的。其中两位工程师,一位是 Frederico Faggin,另一位是 Mashtoshi Shima,曾为 Intel 设计了 4004 和 8080。Zilog[这是一个首字母缩写,Z 表示“the last word”(最后一个字),“i”代表 integrated(集成),“log”代表“logic”(逻辑)]的 Z80 是一个功能非常强大的处理器,包括有片内动态存储刷新电路。这使得系统设计师,例如 Clive Sinclair,只要以很低的成本配很少的外接电路,就可以造成计算机。

一年后,Intel 推出了自己的第一款 16 位处理器,Motorola 也推出了另一种影响很大且经久不衰的 68000 处理器。它可以寻址达 16 MB 的庞大空间。通过片内复杂的电路,它对内可以像 32 位处理器那样运行。这款芯片在 Macintosh、Amiga 和 Atari 等个人计算机中名声大噪。

新理念——RISC

评论家大多将 RISC 视为一个现代理念,更像是一个 1990 年代出现的东西,但实际上它可以追溯到 1965 年 Seymour Cary 的 CDC 6600。RISC 设计强调处理器指令集的简单性,这样就能通过采用先进的结构技术来提高执行这些指令的速度。VAX 的结构就是一个典型的例子。采用了较简单的 VAX 指令后,INDEX 指令的执行速度可以提高 40% 到 60%。CDC6000 具有很多 RISC 的特点,它的指令集很小,只有 64 个操作码,采用加载/存储结构和寄存器到寄存器的操作。此外,指令长 15 位或 30 位,但长度不变。

尽管没有使用 RISC 这个名词,IBM 在 IBM 801(1975 年)——一款发射极耦合逻辑(ECL)多芯片处理器——中正式采用了这些原则。这个设计体系的特点是指令集小,仅有存储器加载/存储操作,24 个寄存器以及采用流水线工作方式。

到了 1980 年代后期 RISC 流行起来以后,IBM 试图将 IBM 801 作为其 OPD 部门(办公产品部)的小型处理器(ROMP)CPU 推向市场,但未获成功。这款芯片后来成为 IBM 3080 的 I/O 处理器的核心。RISC 这一术语被美国两个大学研究项目中的一个首次采用。伯克利(加州大学伯克利分校)的 RISC1 为 SPARK 的商用处理器奠定了基础。与此同时,斯坦福大学的无互锁流水级微处理器(MIPS)已经商业化。

就在斯坦福和伯克利的两个项目的结果发表前后,一家名为 Acorn 的英国家用计算机小公司正在四处寻找新的处理器,取代用于公司当时的计算机系列的 6502。他们对商

用微处理器,包括流行的 8086 和 6800,进行了评估,得出结论认为都不够先进。于是到 1983 年开始自行设计 RISC 微处理器。结果,ARM(高级 RISC 机器公司,前 Acron RISC 机器公司)成为了可能是所有微处理器制造商中真正最接近 RISC 的。

并行处理思想——Transputer

1979 年,英国政府成立了 Inmos 公司,生产新的硅(技术)产品以便在全球舞台上进行竞争。公司的成立,部分原因是为了应对美国对市场日益增强的支配,也是为了满足英国对制造能力的需求。到了 1980 年夏季,Inmos 着手开发自己的第一个芯片。但是事情由于两名工程师对于微处理器的结构的设想各执一词而进展并不顺利。两名工程师,一名是从沃里克大学招聘过来的 David May,另一名是从编写计算机程序复杂的专业公司 Scicon 过来的 Robert Milne。迈恩认为 Transputer——为 Inmos 的处理器起的名字——应是世界上为运行 Ada 软件而量身订做的第一个处理器。他认为这是微处理器设计的未来方向。这和 May 以及 Tony Hoare 意见尖锐对立。Tony Hoare 是来自沃里克大学的学术权威,他曾在那所大学里与 May 共事,两人对 Transputer 的设计有一个共同的简化设计理念。Inmos 的当家人 Iam Barron 无法容忍双方无休止的顶牛,便把自己的观点强加给项目组。他的观点正好符合 May 的想法,同时又提出了多个处理器同时运行的构想。Transputer 于 1985 年以 T-212 的名字推向市场。这个最初的版本是个 16 位的处理器,有类似 RISC 的指令集,每个芯片都有独特的 4 个半行链接(link),使得每个 Transputer 都可以在网络中与其他 Transputer 相连。1994 年,T-9000 发表。这个优化设计采用了脉冲阵列结构,用到并行处理计算机上。

超级 RICS

1988 年,DEC 成立了一个小组来开发新的处理器结构。虽然早在 11 年前,DEC 已从 PDP-11 结构转向了 VAX 结构,但是这时已经可以看出,到 1990 年代早期就会开始落后于人。这个项目结果变得十分庞大,共有 10 个国家的 30 多个工程小组参与到这个后来名为 AlphaAXP 结构体系的设计之中。

这个团队得到了一个极佳的机会去设计一种结构,将 DEC 带入 21 世纪。为了让这个微处理器能有 15 到 25 年的生命期,设计者回顾了过去 25 年的情况,结论是 25 年间计算机能力增加了 1000 倍。他们对未来 25 年做了同样的设想,因此结论是他们设计的产品在未来将以 10 倍的时钟速度、10 倍的指令吞吐率(10 倍超级标量)并用 10 个处理器同时在一个系统中工作。为了使处理器能有效运行多操作系统,他们采用了新的方法,将中断、异常、存储器管理和出错控制纳入一种称为 PAL 指令(特权结构库)的代码中,能够像一般微指令都能做的那样去访问 CPU 硬件。这样一来,Alpha 就堂堂正正地有了一种独特的结构体系。

3.2.5 Reading Materials

A Timeline of Microprocessors (II)

The Eighties

The 1980s, the digital age. This is the time when everything exploded. All the chips we love (and love to hate) were born here—the 286 (possibly Intel’s most crippled(最受委曲的, 最被低估的) chip in its time); the 68020 which was not only a big step forward from the 68000 for its instruction set, but also for being the first 32 bit processor; The ARM CPUs (including the first marketed RISC processor); The 386 and 486 which brought PCs into the 32 bit era; As well as RISC products from Sun (Sparc), MIPS (R3000), and IBM (ROMP).

The beginning of this decade is also when the first clones of Intel CPUs began to appear; NEC brought out the V20 and V30, which are drop-in replacements. The prior Zilog Z80 (1976) executed Intel 8080 instructions but was not a direct replacement for the 8080.

1981: 80186 and 80188 (Intel) x86-compatible, primarily used in embedded systems as they contain DMA and timer circuits.

1982: 80286 (Intel) Used in the IBM PC-AT. (February 1, 1982)

1984: MC68020 (Motorola) The first true 32-bit microprocessor.

1984: V20 and V30 (NEC) First clones of Intel’s 8088 and 8086, respectively.

1985: ARM1 and ARM2 (Acorn) ARM2 was the first commercially available RISC processor. ARM1 is also RISCish, but never made it to market.

1985: R2000 (MIPS) First commercially available MIPS processor.

1986: 80386 (Intel) x86 goes 32 bit.

1986: ROMP (IBM) RISC processor used in the IBM RT PC, a business system which failed in part because of its name (“personal computer”). Successor of IBM 801, predecessor of IBM POWER architecture which eventually lead to PowerPC.

1987: Sparc (Sun and LSI Logic) Sparc is actually a UC Berkeley-derived design for a truly RISC processor, IE one which executes one operation per cycle. The first Sparc CPUs rolled out in 1987 to replace the 68000-family CPUs Sun was previously using to build their systems.

1987: MC68030 (Motorola) 32 bit processor with 32 bit address bus, used in Macintosh, Sun, and Amiga computers (among many others.)

1988: R3000 (MIPS) 32 bit processor, used in many SGI systems, and among other things, the original Sony Playstation.

1988: 80386SX (Intel) Cheaper alternative to the 386DX, it uses a 16 bit time-

multiplexed bus(分时复用总线) to perform 32 bit data transfers (in two cycles) at a cost in memory bandwidth(以减小存储器线宽为代价). (June 16, 1988)

1989: 80486 (Intel) New 32 bit processor, and the last Intel-made x86 processor that is not internally RISC. (April 10, 1989)

1989: ARM3 (Acorn) .

3.3 History of the Development of the ARM Chip at Acorn

3.3.1 Text

Acorn

The first ARM chip, the Acorn RISC Machine (which was later changed and referred to Advanced RISC Machine), was developed by a advanced research and development team at Acorn Computers, a pioneering developer of microcomputers in the UK. At the time Acorn was one of the leading names in the British personal computer market.

Acorn's initial success was sealed when the British Broadcasting Corporation (BBC) commissioned a new home computer model from the company to be sold as the BBC Microcomputer. The release of the BBC Micro in 1982 caught the crest of the home computer wave in Britain, and the BBC name gave Acorn's design the added credibility compared with competing machines from many other developers in this market. The BBC Micro was based around the 8 bit 6502 processor from Rockwell, the same chip that powered the Apple II. Initial models featured colour graphic and 32 Kbytes of random access memory. Data was stored on audiocassettes; hard and floppy disk drive interfaces were also available.

The First ARM

Work on the development of what was to become the ARM began in 1983. Working samples were received in 1985. Steve Furber, now ICL Professor of Computer Engineering at Manchester University, Roger Wilson, who had worked on the design of the BBC Micro, and Robert Heaton who led the VLSI design group within Acorn made up the team developing it. The team worked to create a chip which met their requirements of a processor which retained the ethos of the 6502 but in a 32-bit RISC environment, and implemented this in a small device which it would be possible to design and test easily, and to fabricate cheaply. The important initial decisions were to use a fixed instruction length and a load/store model. Other design decisions were made on an instruction by instruction basis.

The first model of the ARM (ARM1) instruction set was written in BASIC. The subsequent model of the ARM hardware was written in BASIC as well. The actual

physical design of the chip was achieved using VLSI Technology's custom design tools. In addition, an event-driven simulator was designed, also in BASIC, which allowed the support chips, the video controller VIDC and memory controller MEMC, and the I/O controller IOC, to be designed and tested. A development of the simulator, since rewritten in Modula-2 and then in C and known as ASIM, is still used by both Acorn and ARM LTD for design and testing today.

The world's first commercial RISC processor and first ARM processor, ARM1, yield working silicon the first time it was fabricated, in April 1985 at VLSI Technology. It bettered the stated design goals while using fewer than 25,000 transistors. These samples were fabricated using 3 μm process.

Improvements and Developments—ARM2

The experience of designing ARM1, and of programming the sample chips, showed that there were some areas where the instruction set could be improved in order to maximize the performance of systems based around it^[1]. In particular, the Multiply and Multiply and Accumulate instructions were added. The addition facilitated real-time digital signal processing, which was to be used to generate sounds, an important feature of home and educational computers. A coprocessor interface was also added to the ARM at this stage, which would enable a floating point accelerator and other coprocessors to be used with the ARM. The later developed ARM2 still maintained its small die size and low transistor count with all these additions.

ARM, Archimedes, Acorn, and the Market

In 1985 a financial crisis enveloped Acorn and led to it being taken over by the Italian giant Olivetti Ing et Cie, one of Europe's leading computer and office equipment manufactures. The company took over without knowing that Acorn's research labs housed the first sample of a new family of RISC processors. Although the ARM processor had been with the clear intention that it was to power the next generation of Acorn personal computers, and it was equally clear that such a machine needed to be developed quickly, the design and production of ARM-based system by Acorn was to be more fraught than the design of the chip themselves. It was to take more than two years from the arrival of working ARM silicon to the launch and shipment of a complete ARM-based system.

In 1987, a home computer, the Archimedes, was launched as the first commercial using the ARM, featuring an 8 MHz version of the ARM2 and the three support chips MEMC, VIDC, and IOC, an input/output controller and a simple operating system. The Archimedes received a lukewarm response on its launch because personal computing appeared to be consolidating behind the IBM PC standard while Acorn had introduced a

computer with a new processor, a new operating system, and no base of software to provide users with the applications they needed^[2]. It took two to three years for a credible amount of applications software native to the ARM and Archimedes to be developed^[3]. Since then Acorn has refined and improved its computer models and confirmed its position as a leader in the British home and educational computing market.

ARM3 and ARM2aS

After the launch of the Archimedes, Acorn continued to support its research and development team in creating improved versions on the chip, offering greater performance. To expand the design so that it offered the kind of performance expected of a high-end personal computer and or workstation, a 4K byte on-chip data and instruction cache was added. And in 1989 the ARM3 was launched at the significantly increased clock rate of 25 MHz. Acorn's desktop computers using this chip were launch in 1990.

A static version of the processor the ARM2aS was developed soon after the ARM3. This variant added low power consumption to the list of features which made the ARM attractive to developers interested in designing low-cost portable and hand-held devices and electronic personal organizers, and communication device, which although developed as far as working prototypes was never actually marketed^[4].

ARM Ltd and ARM6

Interest in the ARM family was growing as more designers became interested in RISC, and the ARM's design was seen to match a definite need for high-performance, low power consumption, low-cost RISC processors. In conditions of greatest secrecy an agreement was reached between Acorn, VLSI Technology Inc. and a company which had expressed an interest in it for sometime, Apple. The Acorn RISC Machine became the Advance RISC Machine and the company Advances RICS Machines Ltd was born. ARM Ltd. was found with a clear mission to continue the development of the ARM processor and to facilitate its use by system developer, whether as a standalone processor.

ARM Ltd's first development was the next step from the ARM3 processor, which was named ARM6 and included full 32-bit addressing and (endedness support). An improved video controller, VIDC20 was also developed and a floating point processor was also introduced.

The ARM up to Date

As the market for low-cost low power consumption, high performance processors expands, ARM Ltd expanding it global presence by developing relationships with more companies around the world. Since the launch, ARM has developed relationships with

more foundries who have licensed and still license its design and sell them in different markets. From its earliest days with Acorn, ARM Ltd has worked closely with VLSI Technology, Inc. , its first partner and the first manufacturer of ARM devices. In the UK, GEC Plessey Semiconductors was signed as an ARM foundry and partner in January 1992. Plessey now produces a range of ARM standard parts. It is also the foundry for the ARM250. In March 1993 Sharp Corporation of Japan signed a deal to manufacture and market ARM processors and associated products.

ARM Ltd now has offices in California and Japan in order to maintain a close relationship with licensees and their major customers, and to promote existing ARM devices and the company’s ability to produce new ones to future customers. Today ARM continues to establish relationships with new partners like, SUN microsystem and many others, around the world.

At Last the Conclusion

In conclusion, from being a single design aimed at a particular project the ARM is now a set of highly customized processors and supporting micro cells suitable for the use in a wide range of applications but targeted at systems requiring high performance from a compact device with low power consumption.

3.3.2 Specialized English Words

ethos	观念,思想	lukewarm response	反应冷淡
bettered	胜过,超过	creditable	可以认可的,值得赞扬的
working silicon	工作芯片(常指由硅片制成的首个和首批芯片)	native to	原产的,属于……的
ICC Innovative Computing Laboratory		desktop	桌面(计算机),台式(计算机)
(英国曼彻斯特大学)创新计算机实验室		Personal Organizer	一种个人信息管理软件
stated	规定的	foundries	铸造厂(这里泛指工厂)
fraught	令人担忧的		

3.3.3 Notes

[1]The experience of designing ARM1, and of programming the sample chips, showed that there were some areas where the instruction set could be improved in order to maximize the performance of systems based around it. 这个复合句可分为个三层次:主句主语为“The experience of designing ARM1, and of programming the sample chips”,谓语动词为“showed”,“that there were some areas”为其宾语从句;从句中的“areas”又有“where the instruction set could be improved in order to maximize the

performance of systems based around it”做它的定语从句。全句可译为“设计 ARM1 及对样片进行编程的经验表明,指令系统有几处值得改进的地方,以便以指令系统为基础的系统性能得以尽量提高。”

[2]The Archimedes received a lukewarm response on its launch because personal computing appeared to be consolidating behind the IBM PC standard while Acorn had introduced a computer with a new processor, a new operating system, and no base of software to provide users with the applications they needed. 这是一个多重主从复合句。句中“because”以下为从句,其中“while”又引起一转折从句。句末处“they need”为“base of...application”的定语从句。全句可译为“市场对 Archimedes 反应冷淡,因为当时市场似乎是 IBM PC 一统天下,而 Arcon 却推出了一个使用新处理器和新操作系统的计算机,却又没有提供用户应用所需的软件库。”

[3]It took two to three years for a credible amount of applications software native to the ARM and Archimedes to be developed. 句中“native to the ARM and Archimedes”系“application software”的定语。整句结构为一固定表达:“It takes (time period) for (somebody) to (do something)”。此句不定式用了被动语态“to be developed”,与“a creditable amount...”相呼应。全句可译为“大约需要两三年时间,才能开发出根植于 ARM 和 Archimedes 的被用户认同的应用软件。”

[4]This variant added low power consumption to the list of features which made the ARM attractive to developers interested in designing low-cost portable and hand-held devices and electronic personal organizers, and communication device, which although developed as far as working prototypes was never actually marketed. 在这个复合句中,有两个“which”引导的名词性从句,前一从句为限定性定语从句,修饰“low power consumption”;后一从句则是非限定性定语从句,修饰整个主句。其中“although...prototypes”为状语成分,可以认为在“although”之后省去了“it was”。全句可译为“这个品种在一系列特点中又增加了低功耗这一特点,这使得 ARM 对那些致力于设计低价位的便携式和手握式设备、电子个人信息管理系统以及通信设备的开发者很有吸引力。”

3.3.4 Reference Translation

Acron 的 ARM 芯片开发历程

Acron

第一块 ARM 芯片,Acron RISC Machine (这个称谓后来演变成 Advanced RISC Machine),是由英国的一家微型计算机开发商 Acron Computers 的高级研发团队开发出来的。当时的 Arcon 是英国个人计算机市场名列前茅的公司。

Arcon 最初的成功被封存起来了。当年英国广播公司(BBC)委托 Acron 开发一种新型家用计算机,以 BBC 的名义出售。BBC Micro 于 1982 年的发布,可算是抓住了英国家用计算机的鼎盛时期,而 BBC 的显赫大名增加了 Arcon 的设计在市场上和众多其他开发

商的产品竞争时的可信度。BBC Micro 基于 Rockwell 的 8 位处理器 6520——这是组成 Apple II(苹果 II 计算机)的同一款芯片。最初的型号的特点在于具有彩色图形显示及 32 KB 随机访问存储器。数据存储在盒式录音带上,还配有硬盘和软盘驱动器接口。

第一款 ARM

开发演变成成为 ARM 的工作始于 1983 年。1985 年制成样品。现为曼彻斯特大学计算机工程系教授的 Steve Burber,当时从事 BBC Micro 设计工作的 Roger Wilson,还有当年曾领导 Acorn 的 VLSI 设计组的 Robert Heaton,三人组成了开发团队。这个团队致力于开发这样一款处理器,它必须保留 6502 的基本思想,又必须是 32 位 RISC 的构架;还应当是设计调试容易、制造便宜的一款芯片。最初的重要决策包括采用固定长度的指令和冯·诺依曼式加载/存储模式,而其他设计决策则是后来逐条作出的。

第一款 ARM(ARM1)的指令系统样本是用 BASIC 写成的,接下来的一个机型硬件样本也是用 BASIC 写的。芯片的实际设计则采用了 VLST Technology 公司的订制设计工具。此外,还设计了一个硬件驱动的仿真器,也是用 BASIC 写成的,用于对多种支持芯片、视频控制器 VIDC、存储器控制器 MEMC 以及 I/O 控制器 IOC 进行设计和测试。这个仿真器后来又用 Modula-2 以及 C 进行了重写,称为 ASIM,至今还被 Acorn 和 ARM LTD 两家公司用于设计和测试。

世界上第一款商品化 RISC 处理器和第一款 ARM 处理器芯片于 1984 年 4 月在 VLSI Technology 公司制成。它超过了当初规划的设计目标,却用了不到 25 000 只晶体管,采用 3 μ m 制造工艺。

改进与发展——ARM2

设计 ARM1 及对样片进行编程的经验表明,指令系统有几处值得改进的地方,以便以指令系统为基础的系统性能得以尽量提高,尤其是增加了乘和乘-累加指令。这种改进使得实时数字信号的处理变得更加容易。这一功能将用于声音的产生,而发声是家用及教学计算机的重要特点。ARM 还增加了一个协处理器接口。这样,ARM 可以和浮点累加器及其他协处理器一道工作。虽然增加了这些新功能,新开发的 ARM2 仍保持了小尺寸和少晶体管的特点。

ARM、Archimedes、Acorn 与市场

1985 年,Acorn 发生了一场财政危机,导致 Acorn 被欧洲领头的计算机和办公设备制造商之一的意大利巨擘 OliveAi Inget Cie 接手,但 OliveAi Inget Cie 却不知道 Acorn 的研发部门拥有一款新系列的 RISC 处理器样片。虽然 ARM 的意图一直很明确,即为了大力推进 Acorn 新一代的个人计算机,但同样明确的是,这样的计算机需要快速发展,而整个 Acorn 的基于 ARM 的计算机系统的设计和生情况看来比芯片的设计更令人担忧。从完成 ARM 芯片到发布销售完整的 ARM 系统机估计要两年多的时间。

1987 年,一款称为 Archimedes 的家用计算机发布了。它是第一个采用 ARM 的商业化产品,内装 8 MHz 的 ARM2 和三个支持芯片 MEMC、VIDC、IOC 以及一个输入/输出控制器,配上一个简单的操作系统。市场对 Archimedes 反应冷淡,因为当时市场似乎

是 IBM PC 一统天下,而 Arcon 却推出了一个用新处理器和新操作系统的计算机,却又没有提供用户应用所需的软件库。大约需要两三年时间,才能开发出根植于 ARM 和 Archimedes 的被用户认同的应用软件。此后,Acron 不断更新改进它的计算机系统,并证明自己在英国家庭和教育计算机市场的领军地位。

ARM3 和 ARM2aS

Archimedes 发布后,Acorn 继续支持它的研发团队开发 Archimedes 的更新版本,提供更强的性能。为了扩大设计,以便为高端个人机和工作站提供所预期的性能,芯片增加了 4 KB 的片内数据及指令缓存器。到 1989 年,ARM3 发布了。它的时钟频率大大提高了,达到 25 MHz。采用这款芯片的 Arcon 台式机于 1990 年发布。

ARM3 之后不久,发布了 ARM 的静态版本 ARM2aS。这个品种在一系列特点中又增加了低功耗这一特点,这使得 ARM 对那些致力于设计低价位的便携式和手握式设备、电子个人信息管理系统以及通信设备的开发者很有吸引力。尽管已经开发出可用的原型产品,但 ARM2aS 从未真正进入市场。

ARM 公司和 ARM6

随着更多的设计人员对 RISC 发生兴趣,人们对 ARM 系列的兴趣也在增长。人们看到 ARM 的设计明确满足了对高性能、低功耗、低价格的 RISC 处理器的需要。在最大限度保密的情况下,Acron、VLSI Technology Inc 和 Apple——Apple 一段时间以来就对此事表示有兴趣——达成了一项协议,将 Acorn RISC Machine 变更为 Advance RISC Machine,这样就诞生了 Advance RISC Machine 有限公司。ARM 公司的明确目标就是继续开发 ARM 处理器,使其可以为系统开发者所使用,不管 ARM 自身是否能成为一个独立的处理器。

ARM 公司的第一项开发目标就是从 ARM3 走向下一步——ARM6,它能全 32 位寻址,并改进了视频控制器 VIDC20,引入了一个浮点处理器。

最新 ARM

随着低价位、低功耗、高性能处理器市场的扩展,ARM 公司通过和全球更多的公司发展关系,扩大了自己在世界的影响。ARM 问世以来,ARM 公司就和多家制造厂发展了关系,这些工厂拥有 ARM 设计的生产许可,在不同的市场上出售产品。从最早的 Acorn 开始,ARM 公司就和作为它的第一个合伙人以及 ARM 产品生产商的 VLSI Technology 公司密切合作。在英国,GEC Plessey Semiconductors 公司于 1992 年 1 月作为生产厂家和合伙人和 ARM 公司签约。Plessey 如今生产一系列 ARM 标准产品,它也是 ARM250 的生产厂家。1993 年 3 月,Sharp Corporation of Japan(夏普)也和 ARM 签订了协议,生产销售 ARM 处理器和相关产品。

ARM 公司现在在加州和日本均设有办事机构,以便和获得生产许可的厂家以及重要用户保持密切关系,推进已有的 ARM 产品的销售,扩大公司为未来用户生产新产品的能力。如今,ARM 继续和新的合作伙伴建立关系,例如 SUN 公司和世界各地的很多其他公司。

最后的结论

总而言之,从最初针对某一具体目标的单项设计发展至今,ARM 已成为一系列面向体积紧凑、低功耗的高性能系统的订制处理器和支持性微构件,应用领域十分广阔。

3.3.5 Reading Materials

The ARM Cortex-M3 Processor

Higher Performance Through Better Efficiency

In order to achieve higher performance, processors can either work hard or work smart. Pushing higher clock frequencies may increase performance but is also accompanied by higher power consumption and design complexity. On the other hand, higher compute efficiency at slower clock speeds results in simpler and lower power designs that can perform the same tasks. At the heart of the Cortex-M3 processor is an advanced 3-stage pipeline core(三级流水线内核), based on the Harvard architecture, that incorporates many new powerful features such as branch speculation(转移预测), single cycle multiply and hardware divide to deliver an exceptional Dhrystone benchmark performance of 1.25 DMIPS/MHz. The Cortex-M3 processor also implements the new Thumb®-2 instruction set architecture, helping it to be 70% more efficient per MHz than an ARM7TDMI-S processor executing Thumb instructions, and 35% more efficient than the ARM7TDMI-S processor executing ARM instructions, for the Dhrystone benchmark.

Ease of Use for Quick and Efficient Application Development

Reducing time-to-market(开发周期, 上市时间) and lowering development costs are critical criteria in the choice of microcontrollers, and the ability to quickly and easily develop software is key to these requirements.

The Cortex-M3 processor has been designed to be fast and easy to program, with the users not required to write any assembler code(汇编程序) or have deep knowledge of the architecture to create simple applications. The processor has a simplified stack-based programmer's model which still maintains compatibility with the traditional ARM architecture but is analogous to the systems employed by legacy 8-bit and 16-bit architectures, making the transition to 32-bit easier. Additionally a hardware based interrupt scheme means that writing interrupt service routines (handlers) becomes trivial(无足轻重), and that start-up code is now significantly simplified as no assembler code register manipulation is required.

3.4 Memory Organization in MCS-51 Family of Microcontrollers

3.4.1 Text

Logical Separation of Program and Data Memory

All MCS-51 devices have separate address spaces for Program and Data Memory, as shown in Figure 3.4.1. The logical separation of Program and Data Memory allows the Data Memory to be accessed by 8-bit addresses, which can be more quickly stored and manipulated by an 8-bit CPU^[1]. Nevertheless, 16-bit Data Memory addresses can be generated through the DPTR register.

Program Memory can only be read, not written to. There can be up to 64K-bytes of Program Memory. In the 8051, 8051AH 80051BH, and their EPROM versions, the lowest 4K bytes of Program Memory are on-chip. The 8052AH provides 8K-bytes of on-chip Program Memory storage. In the ROMless versions (8031, 8031AH, 80C31BH, 8032AH) all Program Memory is external. The read strobe for external Program Memory is the signal $\overline{\text{PSEN}}$ (Program Store Enable).

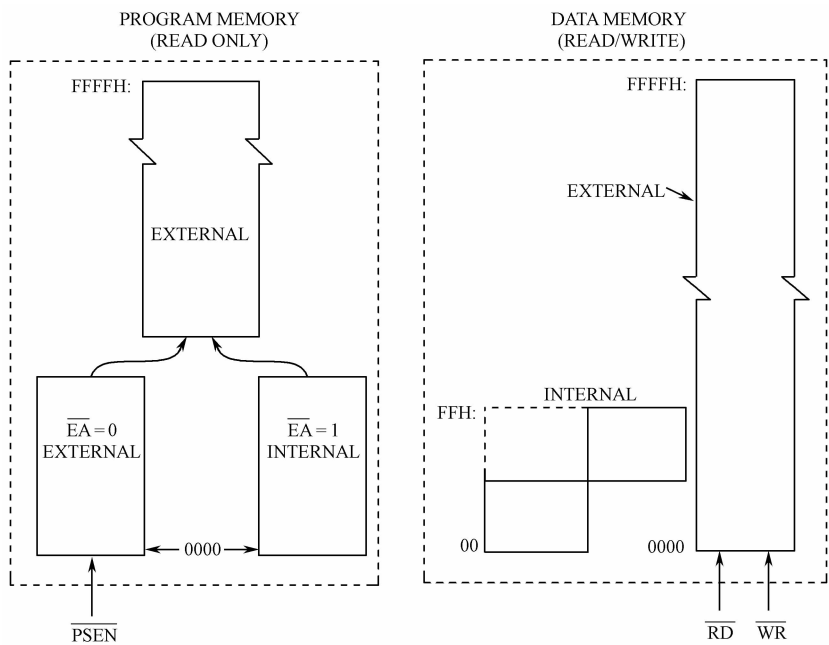


Figure 3.4.1 MCS-51 Memory structure.

Data Memory occupies a separate address space from Program Memory. Up to 64K-bytes of external RAM can be addressed in the external Data Memory space. The CPU

generates read and write signals, \overline{RD} and \overline{WR} , as needed during external Data Memory accesses.

External Program Memory and external Data Memory may be combined if desired by applying the \overline{RD} and \overline{PSEN} signals to the inputs of an AND gate and using the output of gate as the read strobe to the external Program/Data Memory^[2].

Program Memory

Figure 3. 4. 2 shows a map of the lower part of the Program Memory. After reset, the CPU begins execution from location 0000H.

As shown in Figure 3. 4. 2, each interrupt is assigned a fixed location in Program Memory. The interrupt causes the CPU to jump to that location, where it commences execution of the service routine^[3].

External Interrupt 0, for example, is assigned to location 0003H. If External Interrupt 0 is going to be used, its service routine must begin at location 0003H.

If the interrupt is not going to be used, its service location is available as general purpose Program Memory.

The interrupt service locations are spaced at 8-byte intervals: 0003H for External Interrupts 0, 000BH for Timer 0, 0013H for External Interrupt 1, 001BH for Timer 1, etc. If an interrupt service routine is short enough (as is often the case in control applications), it can reside entirely within that 8-byte interval. Longer service routines can use a jump instruction to skip over subsequent interrupt locations, if other interrupts are in use^[4].

The lowest 4K (or 8K, in the 8052AH) bytes of Program Memory can be either in the on-chip ROM or in an external ROM. This selection is made by strapping the \overline{EA} (External Access) pin to either V_{CC} or V_{SS} .

In the 8051 and its derivatives, if the \overline{EA} pin is strapped to V_{CC} , then program fetches to addresses 0000H through 0FFFH are directed to the internal ROM. Program fetches to addresses 1000H through FFFFH are directed to external ROM.

In the 8052AH, $\overline{EA} = V_{CC}$ selects addresses 0000H through IFFFH to be internal, and addresses 2000H through FFFFH to be external.

If the \overline{EA} pin is strapped to V_{SS} , then all program fetches are directed to external ROM. The ROM less parts (8031, 8032AH, etc.) must have this pin externally strapped to V_{SS} to enable them to execute from external Program Memory.

The read strobe to external ROM, \overline{PSEN} is used for all external program fetches. \overline{PSEN} is not activated for internal program fetches.

The hardware configuration for external program execution is shown in

Figure 3.4.3. Note that 16 I/O lines (Ports 0 and 2) are dedicated to bus functions during external Program Memory fetches. Port 0 (P0 in Figure 3.4.3) serves as a multiplexed address/data bus. It emits the low byte of the Program Counter (PCL) as an address, and then goes into a float state awaiting the arrival of the code byte from the Program Memory. During the time that the low byte of the Program Counter is valid on P0, the signal ALE (Address Latch Enable) clocks this byte into an address latch^[5]. Meanwhile, port 2 (P2 in Figure 3.4.3) emits the high byte of the Program Counter (PCH). Then $\overline{\text{PSEN}}$ strobes the EPROM and the code byte is read into the microcontroller.

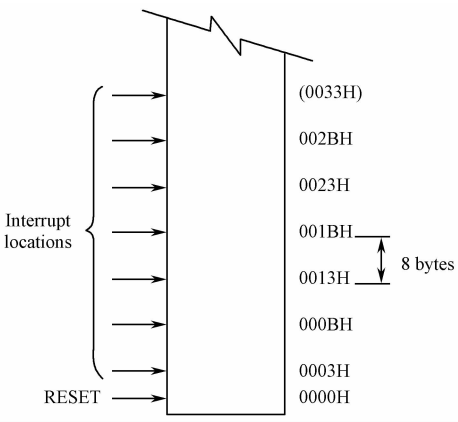


Figure 3.4.2 MCS-51 Program Memory.

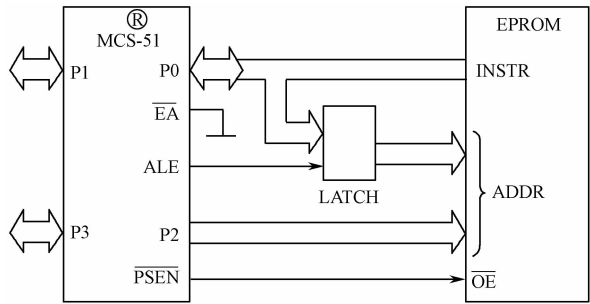


Figure 3.4.3 Executing from external Program Memory.

Program Memory addresses are always 16 bits wide, even though the actual amount of Program Memory used may be less than 64K bytes. External Program execution sacrifices two of 8-bit ports, P0 and P2, to the function of addressing the Program Memory.

Data Memory

The right half of Figure 3.4.1 shows the internal and external Data Memory spaces available to the MCS-51 user.

Figure 3.4.4 shows a hardware configuration for accessing up to 2K bytes of external RAM. The CPU in this case is executing from internal ROM. Port 0 serves as a multiplexed address/data bus to the RAM, and 3 lines of port 2 are being used to page the RAM. The CPU generates $\overline{\text{RD}}$ and $\overline{\text{WR}}$ signals as needed during external RAM accesses.

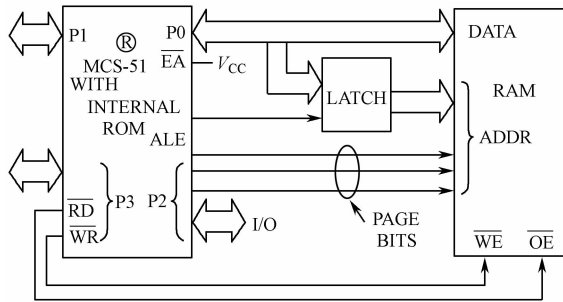


Figure 3.4.4 Accessing external Data Memory. If the Program Memory is internal, the other bits of P2 are available as I/O.

There can be up to 64K bytes of external Data Memory. External Data Memory addresses can be either 1 or 2 bytes wide. One-byte addressed are often used in conjunction with one or more other I/O lines to page the RAM, as shown in Figure 3.4.4^[6]. Two-byte addresses can also be used, in which case the high address byte is emitted at port 2.

Internal Data Memory is mapped in Figure 3.4.5. The memory space is shown divided into three blocks, which are generally referred to as the Lower 128, the Upper 128 and SFR space. Internal Data Memory addresses are always one byte wide, which implies an address space of only 256 bytes. However, the addressing modes for internal RAM can in fact accommodate 384 bytes, using a simple trick^[7]. Direct addresses higher than 7FH access one memory space and indirect addressing higher than 7FH access a different memory space. Thus Figure 3.4.5 shows the Upper 128 and SFR space occupying the same block of addresses, 80H through FFH, although they are physically separate entities.

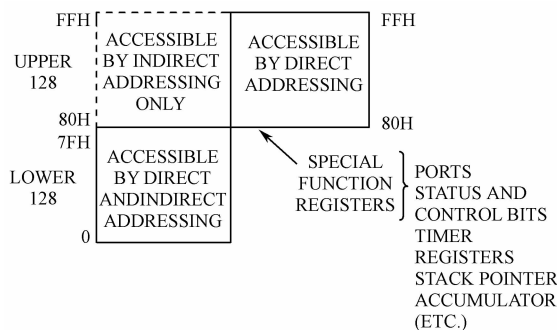


Figure 3.4.5 Internal Data Memory.

The Lower 128 bytes of RAM are present in all MCS-51 devices as mapped in Figure 3.4.6. The lowest 32 bytes are grouped into 4 banks of 8 registers. Program instructions call out these registers as R0 through R7. Two bits in the Program Status Word (PSW) select which register bank is in use. This allows more efficient use of code space, since register instructions are shorter than instructions that use direct addressing.

The next 16 bytes above the register banks form a block of bit-addressable memory space. The MCS-51 instruction set includes a wide selection of single-bit instructions, and the 128 bits in this area can be directly addressed by these instructions. The bit addresses in this area 00H through 7FH.

All of the bytes in the Lower 128 can be accessed by either direct or indirect addressing. The Upper 128 (Figure 3.4.7) can only be accessed by indirect addressing. The Upper 128 bytes of RAM are not implemented in the 8051, but are in the 8052AH.

Figure 3.4.8 gives a brief look at the Special Function Register (SFR) space. SFRs include the port latches, timers, peripheral controls, etc. These registers can only be accessed by direct addressing. In general, all MCS-51 microcontrollers have the same SFRs as the 8501, and at the same addresses in SFR space. However, enhancements to the 8501 have additional SFRs that are not present in the 8501, nor perhaps in other proliferations of the family^[8].

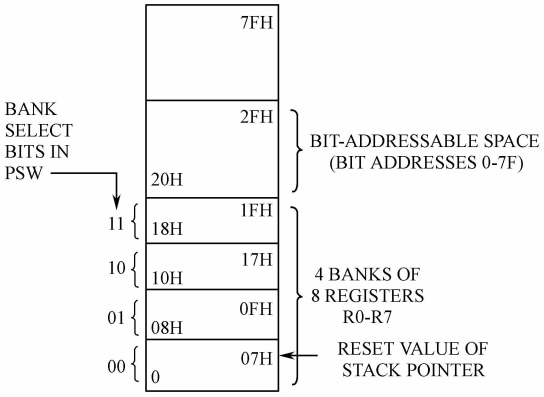


Figure 3.4.6 The lower 128 bytes of internal RAM.

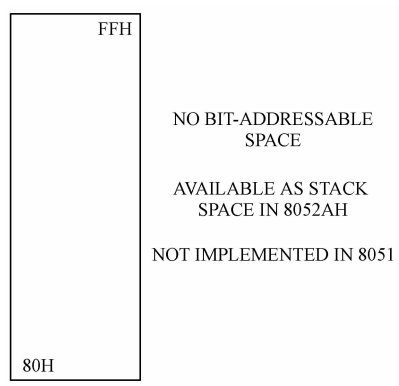


Figure 3. 4. 7 The upper 128 bytes of internal RAM.

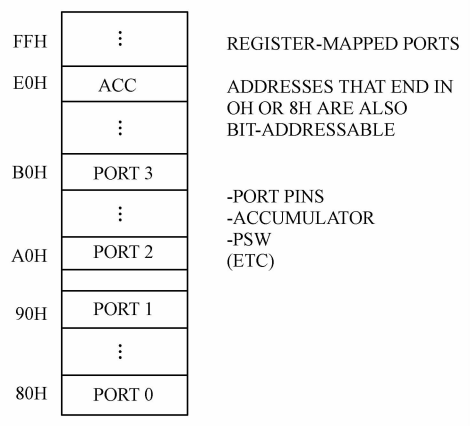


Figure 3. 4. 8 SPR space.

3. 4. 2 Specialized English Words

MCS-51 Family MCS-51 系列
microcontroller 微控制器
address space 地址空间
access 访问
Program and Data Memory 程序存储器和数据存储器
External Access 外部访问(信号), 访问(信号)
derivatives 派生品种, 衍生品种
strap...to... 将.....搭接到.....上
fetch 读取
manipulate 控制
CPU 中央处理器 (Central Processing Unit 之缩写)
emit 发射, 发出
float state 浮置态
Program Counter 程序计数器
code byte 代码字节
clock 计时, 记录
latch 锁存器
ALE (Address Latch Enable) 地址

锁存允许(信号)
addressing 寻址
page 页面, 寻址定位
hardware configuration 硬件配置, 组合
register 寄存器
muhplexed 多路的, 多用的
reside 驻留, 保存
multiplexed address/data bus 地址/数据复用总线
strobe 选通(信号)
Program Store Enable 程序存储器允许(信号)
RAM 随机访问存储器, 读写存储器 (Random Access Memory 之缩写)
lower part (存储器的)低端
map 用图形表示
interrupt 中断
routine 程序
interrupt service routine 中断服务子程序

3.4.3 Notes

[1] The logical separation of Program and Data Memory allows the Data Memory to be accessed by 8-bit addresses, which can be more quickly stored and manipulated by an 8-bit CPU. 句中的“to be accessed”为宾语“the Data Memory”的补语,“which …”则是它的定语,全句可译为“程序存储器和数据存储器的逻辑分离,使得数据存储器可以用八位地址进行访问,这样就可以用一个 8 位的 CPU 对它进行更快的存储与控制。”

[2] External Program Memory and external Data Memory may be combined if desired by applying the \overline{RD} and \overline{PSEN} signals to the inputs of an AND gate and using the output of gate as the read strobe to the external Program / Data Memory. 句中“if”引起条件状语,可以认为其后省略了“it is”。“by …”为一个很长的介词短语,也做状语。此句可译为“如果有必要,可以将外部程序存储器和外部数据存储器合用同一个存储器,方法是将 \overline{RD} 信号和 \overline{PSEN} 信号用做一个‘与门’的输入,而将该与门的输出作为外部程序/数据存储器的读选通信号。”

[3] The interrupt causes the CPU to jump to that location, where it commences execution of the service routine. 句中的“where”引导定语从句,修饰“that location”。从句中的“it”指“CPU”。全句可译为“中断使 CPU 跳转到该地址,从此处开始执行(中断)服务子程序。”

[4] Longer service routines can use a jump instruction to skip over subsequent interrupt locations, if other interrupts are in use. 全句可译为“对于较长的(中断)服务子程序,如果还有其他中断也处在开放状态下,可以用一条转移指令跳过后面的中断地址区。”

[5] During the time that the low byte of the Program Counter is valid on P0, the signal ALE (Address Latch Enable) clocks this byte into an address P0. 句中的“During the time”为介词短语,做状语,而“that the …P0”为“time”的定语。“the signal ALE”为全句的主语,“clock”为谓语动词。全句可译为“当程序计数器的低八位字节在 P0 端口上有效时,ALE 信号将该字节锁存到地址锁存器中。”

[6] One-byte addressed are often used in conjunction with one or more other I/O lines to page the RAM, as shown in Figure 3.4.4. 此句可译为“一字节地址常与其他一位或几位 I/O 线组合共同对 RAM 寻址,如图 3.4.4 所示。”

[7] However, the addressing modes for internal RAM can in fact accommodate 384 bytes, using a simple trick. “using a simple trick”为现在分词短语做状语。此句可译为“用一个简单的小技巧,便可让内部 RAM 的寻址方式实际上可访问 384 个字节。”

[8] However, enhancements to the 8501 have additional SFRs that are not present in the 8501, nor perhaps in other proliferations of the family. 句中 that 引导的定语从句修饰“SFRs”。全句可译为“然而,8501 的增强型品种另外增加了一些 SFR 寄存器,这些寄存器是 8501 所没有的,也许也是 51 系列其他派生品种所没有的。”

3.4.4 Reference Translation

MCS-51 系列微控制器存储器结构

程序存储器和数据存储器的逻辑分离

如图 3.4.1 所示,所有 MSC-51 产品的程序存储器和数据存储器的地址空间都是分开的。程序存储器和数据存储器的逻辑分离,使得数据存储器可以用八位地址进行访问,这样就可以用一个 8 位的 CPU 对它进行更快的存储与控制。然而,可以通过 DPTR 寄存器建立 16 位数据存储器的地址。

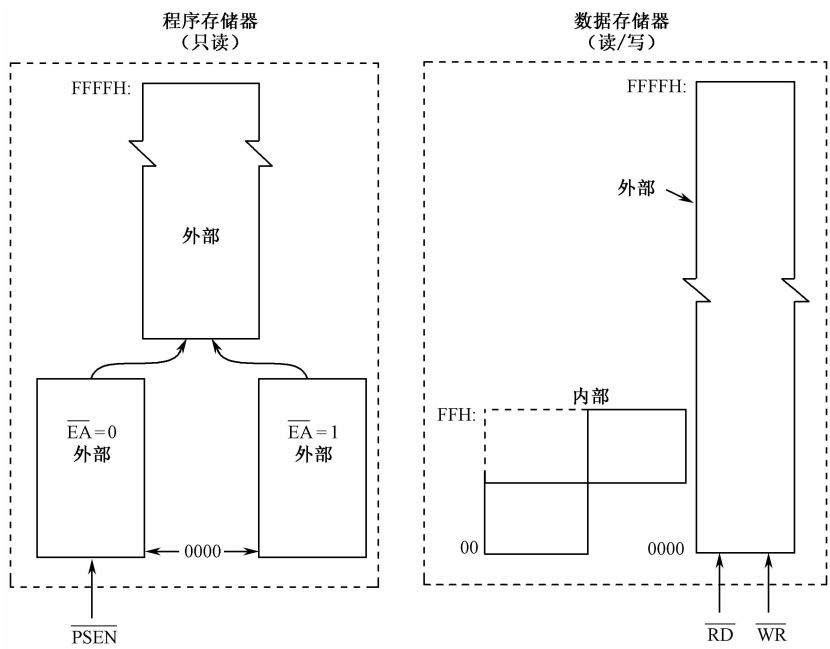


图 3.4.1 MSC-51 存储器结构

程序存储器只能读出,不能写入。程序存储器可以多达 64 KB。在 8051、805AH、80C51BH 以及它们的 EPROM 版本中,最低 4 KB 的程序存储器已制作在片内。8052AH 则有 8 KB 片内程序存储器。在无 ROM 版本(8031、8031AH、80C31BH、8032AH)中,所有的程序存储器都是外部的。外部程序存储器读选通信号是PSEN(程序存储器允许)。

数据存储器占据和程序存储器不一样的地址空间。外部数据存储器可以有多达64 KB 的外部 RAM。访问外部数据存储器时,CPU 根据需要产生读和写信号,RD和WR。

如果有必要,可以将外部程序存储器和外部数据存储器合用同一个存储器,方法是将RD信号和PSEN信号用做一个“与门”的输入,而将该与门的输出作为外部程序/数据存储器的读选通信号。

程序存储器

图 3.4.2 显示的是程序存储器的低端。复位后,CPU 从 0000H 地址开始运行。

如图 3.4.2 所示,每个中断在程序存储器中都分配有一个固定的地址。中断使 CPU 跳转到该地址,从此处开始执行(中断)服务子程序。例如配给外部中断 0 的地址是 0003H。如果打算使用外部中断 0,它的服务子程序必须从地址 0003H 开始存放。如果不打算用外部中断 0,它的服务地址可以用做程序存储器的一般单元使用。

每个中断程序入口地址相隔 8 个字节: 0003H 用于外部中断 0;000BH 用于定时器 0;0013H 用于外部中断 1,等等。对于较长的(中断)服务子程序,如果还有其他中断也处在开放状态下,可以用一条转移指令跳过后面的中断地址区。

存储存储器最低端的 4 KB(8052AH 为 8 KB)可以是片内 ROM,也可以是外部 ROM。这是通过将EA(访外)引脚连接到 V_{CC}或 V_{SS}来进行选择的。

在 8051 及其衍生品种中,如果EA脚连接到 V_{CC},则程序对 0000H 到 0FFFH 的取指操作指向片内 ROM。程序对 1000H 到 FFFFH 的取指操作指向外部 ROM。

在 8052AH 中,EA = V_{CC}时,表示地址 0000H 到 1FFFH 为片内的,地址 2000H 到 FFFFH 为片外的。

如果EA脚接到 V_{SS}上,所有程序取指操作均指向外部 ROM,对于无 ROM 的品种(8031、8032AH 等)而言,必须将EA接至 V_{SS},使其从外部程序存储器执行程序。

外部 ROM 的读选通信号PSEN用于对所有外部程序的取指操作。片内的取指操作不会激活PSEN。

执行外部程序的硬件配置由图 3.4.3 中给出。注意,16 条 I/O 线(端口 0 和端口 2)在访问外部程序存储器时是专门起总线作用的。端口 0(图 3.4.3 中的 P0)用作地址/数据复用总线。它首先发出程序计数器低位字节(PCL)地址,然后转为浮置态,等待来自程序存储器的指令代码字节的到来。当程序计数器的低八位字节在 P0 端口上有效时,ALE 信号将该字节锁存到地址锁存器中。与此同时,端口 2(图 3.4.3 中的 P2)发出程序计数器的高位字节(PCH)。随后PSEN选通 EPROM,指令代码便被读入微控制器。

程序存储器的地址总是 16 位宽,即使程序存储器的实际大小可能会小于 64 KB 也是如此。执行外部程序需要占用两个 8 位端口 P0 和 P2 来实现程序存储器的访问。

数据存储

图 3.4.1 右半部分显示的是 MCS-51 用户可以使用的内部及外部数据存储器空间。

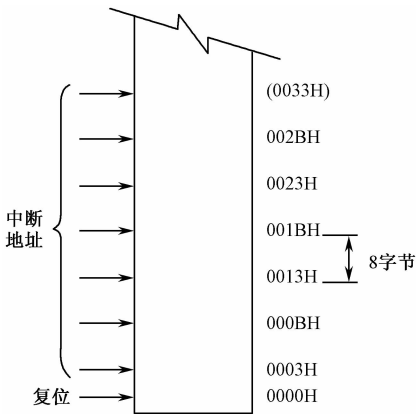


图 3.4.2 MCS-51 程序存储器

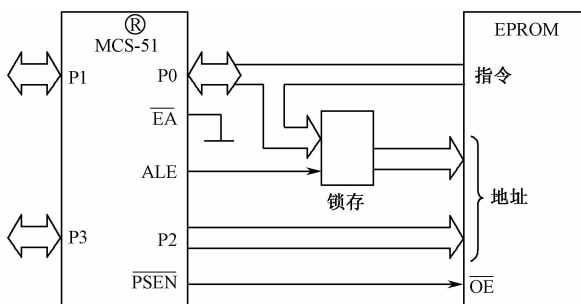


图 3.4.3 从外部程序存储器运行程序

图 3.4.4 给出的是访问最多达 2 KB 外部 RAM 的硬件配置。这时 CPU 从片内 ROM 执行程序。端口 0 对于 RAM 起地址/数据总线的双重作用。端口 2 的 3 条引线用于 RAM 寻址。CPU 根据访问外部 RAM 的实际情况产生 \overline{RD} 和 \overline{WR} 信号。

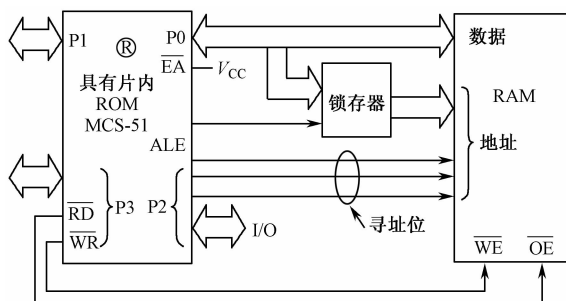


图 3.4.4 访问外部数据存储器。设程序存储器在片内, P2 其余的线可用于 I/O

最多可接入 64 KB 外部数据存储器。外部数据存储器的地址的宽度可以是 1 个字节, 也可以是 2 个字节。一字节地址常与其他一位或几位 I/O 线组合共同对 RAM 寻址, 如图 3.4.4 所示。也可以使用 2 字节地址, 这时高地址字节由端口 2 送出。

图 3.4.5 画出的是内部数据存储器的结构。可以看出存储器空间被分成三块, 一般

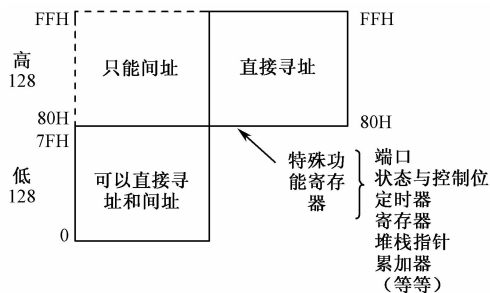


图 3.4.5 片内数据存储器

称为低 128 高 128 和 SFR 空间。用一个简单的小技巧, 便可让内部 RAM 的寻址方式实际上可访问 384 个字节。地址大于 7FH 的直接寻址所访问的是一个存储器空间; 地址大于 7FH 的间址所访问的是另一个存储器空间。所以尽管物理实体不同, 高 128 空间和 SFR 空间在图 3.4.5 中占有相同的地址区域 —— 80H 到 FFH。

如图 3.4.6 所示, RAM 的低 128 字节在所有 MCS-51 机型中都有。其最低的 32

个字节分为 8 个寄存器一组,共 4 组。程序指令称这些寄存器为 R0 至 R7。程序状态字 (PSW)中有两位用于选择使用哪一个组。这样做能更有效地利用代码空间,因为寄存器 (寻址)指令比直接寻址指令短一些。

和寄存器组相邻的 16 个字节构成了位寻址存储空间。MCS-51 指令系统中有大量针对一位进行操作的指令,这些指令可以对这个空间的 128 位进行直接寻址。这个空间的位地址为 00H 到 7FH。

所有低端 128 字节均可直接寻址或间址,而高端 128(见图 3.4.7)只能间址访问。8051 没有高 128 字节 RAM,但 8052HA 有。

图 3.4.8 给出了特殊功能寄存器 (SFR)空间的简图。SFR 包括端口锁存、定时器、外围控制信号等。这些寄存器只能用直接寻址方式访问。一般而言,所有 MCS-51 微控制器都有和 8051 相同的 SFR,它们在 SFR 空间中的地址也相同。然而,8051 的增强型品种另外增加了一些 SFR 寄存器,这些寄存器是 8051 所没有的,也许也是 51 系列其他派生品种所没有的。

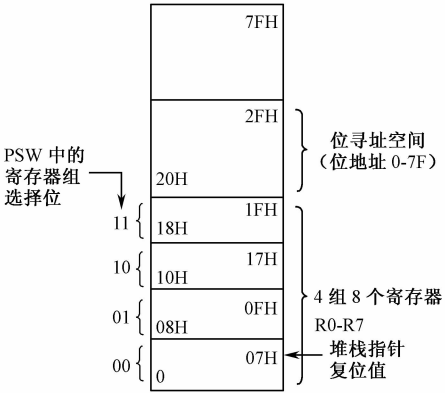


图 3.4.6 片内 RAM 低端 128 字节

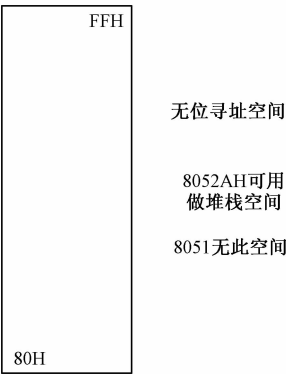


图 3.4.7 片内 RAM 高端 128 字节

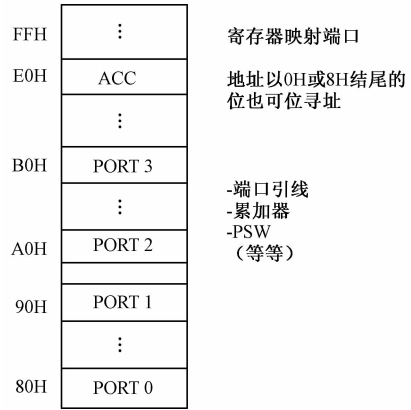


图 3.4.8 SFR 空间

3.4.5 Reading Materials

Standard Serial Interface of MCS-8051

The serial port is full duplex (全双工), meaning it can transmit and receive simultaneously. It is also receive-buffered (接收缓存的), meaning it can commence reception of a second byte before a previously received byte has been read from the

register. (However, if the first byte still hasn't been read by the time reception of the second byte is complete, one of the bytes will be lost.) The serial port receive and transmit registers are both accessed at Special Function Register SBUF. Writing to SBUF loads the transmit register, and reading SBUF accesses a physically separate receive register.

The serial port can operate in 4 modes;

Mode 0: Serial data enters and exits through RxD. TxD outputs the shift clock. 8 bits are transmitted/received (LSB first).

The baud rate is fixed at $1/12$ the oscillator frequency(振荡频率).

Mode 1: 10 bits are transmitted (through TxD) or received (through RxD): a start bit (0), 8 data bits (LSB first), and a stop bit (1). On receive, the stop bit goes into RB8 in Special Function Register SCON. The baud rate is variable.

Mode 2: 11 bits are transmitted (through TxD) or received (through RxD): start bit (0), 8 data bits (LSB first), a programmable 9th data bit, and a stop bit (1). On Transmit, the 9th data bit (TB8 in SCON) can be assigned the value of 0 or 1. Or, for example, the parity bit (P, in the PSW) could be moved into TB8. On receive, the 9th data bit goes into RB8 in Special Function Register SCON, while the stop bit is ignored. The baud rate is programmable to either $1/32$ or $1/64$ the oscillator frequency.

Mode 3: 11 bits are transmitted (through TxD) or received (through RxD): a start bit (0), 8 data bits (LSB first), a programmable 9th data bit, and a stop bit (1). In fact, Mode 3 is the same as Mode 2 in all respects except baud rate. The baud rate in Mode 3 is variable.

In all four modes, transmission is initiated(启动, 开始) by any instruction that uses SBUF as a destination register(目的地寄存器). Reception is initiated in Mode 0 by the condition $RI = 0$ and $REN = 1$. Reception is initiated in the other modes by the incoming start bit if $REN = 1$.

3.5 The Development of Computer-Based Control Systems

3.5.1 Text

The history of industrial automation is a long one. Since the 1930s major progress has been achieved through the existence of both user needs and technological advances. Prior to the introduction of the use of computers for industrial control application, the standard industrial control system, those block diagram is shown in Figure 3.5.1, was that of a large number of single-loop controllers, either pneumatic or electronic. Note that in the Figure 3.5.1, Level 1A refers to those instruments which have no signal read-out readily available to the operator, such as some field-mounted controllers, ratio

instruments, etc^[1]. Level 1B devices have a regular indication to the operator, such as a strip chart, pointer etc. These later are the standard PID (proportional, integral and derivative control mode) controllers.

The first application of digital computers to industrial process was one developed for the plant monitoring of an electric power generation station at Sterlington, Louisiana, plant of Louisiana Power and Light Company in September, 1958, as shown in Figure 3. 5. 2^[2]. Shortly thereafter in March 1959 and April 1960, the first computer supervisory control systems were installed at a refinery in Texas by the TEXACO Company and at an ammonia plant in Louisiana by the Monsanto Chemical Company. Their block diagram is shown in Figure 3. 5. 3.

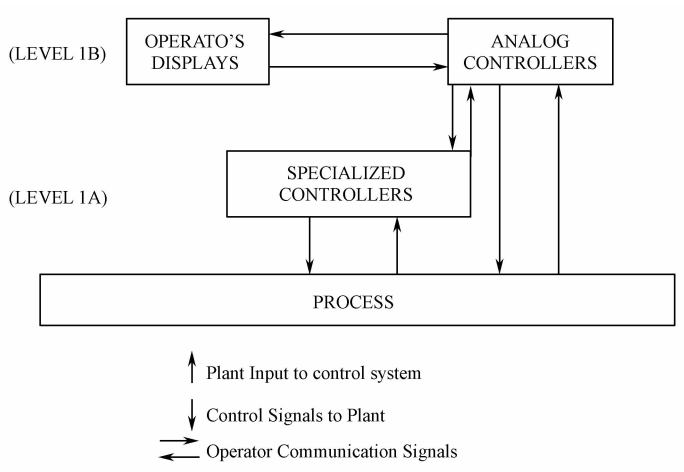


Figure 3. 5. 1 The basic system-analog control.

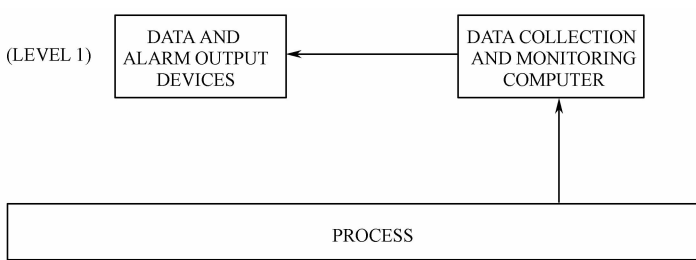


Figure 3. 5. 2 Data collection and monitoring system.

In this kind of computer control system, the computer calculated the control set points and then sent them to the analog controllers instead of controlling the process directly. The analog controllers remained the primary means of process control.

The next logical step in the development of computer control systems was the use of

a computer to bypass the analog controllers of Figure 3. 5. 3 and to control the process directly. This is the so-called direct digital control (DDC) which was achieved for the first time by the Monsanto Company in their ethylene plant at Texas City, Texas in March 1962, as diagrammed in Figure 3. 5. 4. Note that Figure 3. 5. 4 includes the possibility of specialized dedicated digital controllers^[3] to show the relationship to Figure 3. 5. 1. While common today, these devices were not present in the very early DDC systems^[4]. All work was carried out then in the digital computer of Level 1B.

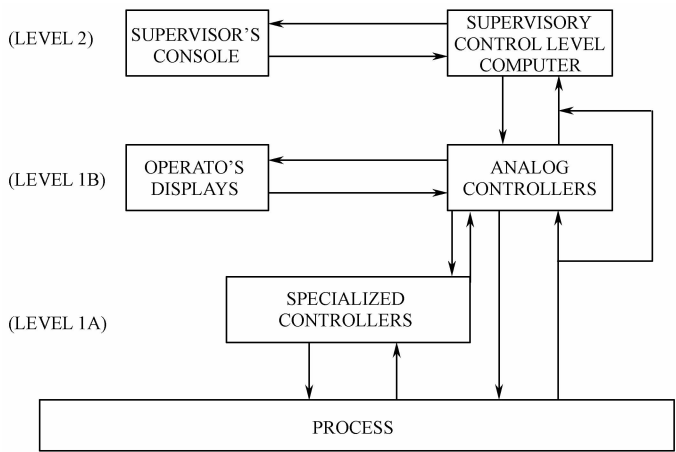


Figure 3. 5. 3 Secondary or supervisory digital control only.

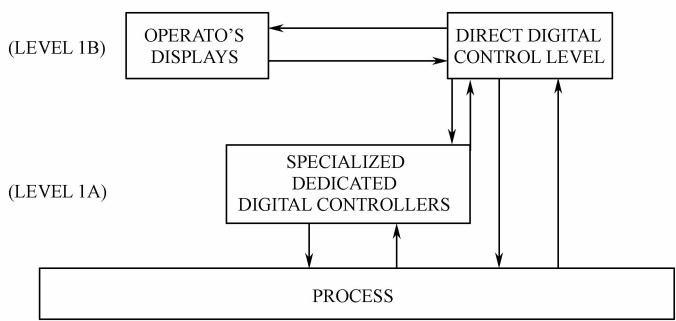


Figure 3. 5. 4 Primary or digital control only.

People should keep in mind that the first development of the digital computer itself started in 1946, only twelve years earlier than it was first used in industrial control systems. Although its very high capabilities and potentialities for engineering applications were widely recognized, and the development of computer science and technology was rapid during the 1950s and 1960s, it was even then (mid 1960s) still in its primary stage. The inherently severe performance requirements imposed upon earlier

digital computers made it necessary for the manufacturers to develop much larger, much faster, and so, much more expensive computers to be built and used^[5]. To help justify the higher cost, the vendors advocated the incorporation of all types of computer functions, including both supervisory control and DDC, in one computer box or mainframe at a central control-room as shown in Figure 3. 5. 5 and Figure 3. 5. 6. We will call this the centralized computer system.

While the centralized computer control system provided significant advantages over the earlier ones, they also suffered from a number of disadvantages. The biggest disadvantage of this architecture was that any failure of the computer itself could unavoidably shut down the entire system and the reliability of the computers was rather low at that time^[6]. Another structural problem was that, because of the centralized computer location, the vast plant communication system required to bring the plant signals to the computer and return control signals to the field was also very expensive and was prone to the electrical noise problem^[7].

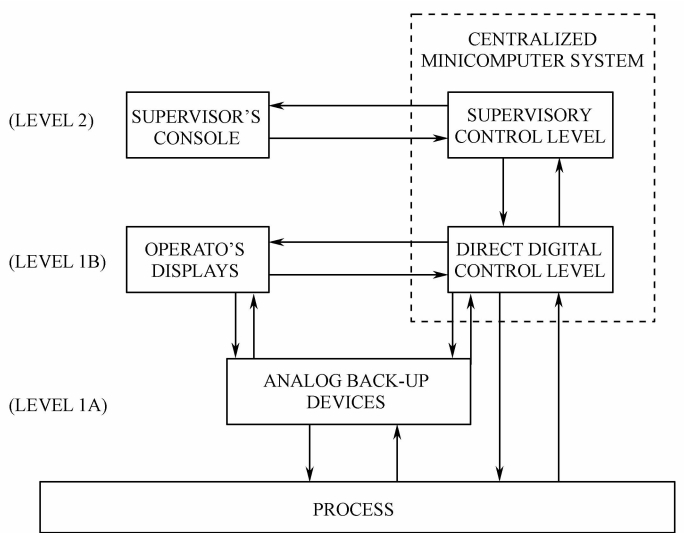


Figure 3. 5. 5 Complete secondary digital control, supervisory plus direct digital control with associated analog control back-up computation provided by a centralized minicomputer.

To compensate for the unreliability of the computer hardware, another “hot standby” computer or a complete analog backup system paralleling the DDC was needed, again greatly increasing its cost^[8].

Again, to compensate these high costs, users and vendors alike attempted to squeeze as large as possible a project into the computer, further worsening the disadvantage of centralization and extremely, complicating its programming by requiring

the computer to perform a wide variety of functions in real time: input scanning, database updating, control algorithm computation, data logging and man-machine interfacing, etc^[9].

The result of these difficulties was a disenchantment by factory management toward computers and an effective hiatus on computer system installations until about the middle of the 1970s.

It had become clear that a new approach was necessary to assure the capabilities in speed, reliability and memory capacity needed to fulfill the needs and expectations of the users. Perhaps a new architectural approach was needed—the power of the computer system should be scattered about the plant to avoid the limitations imposed by using one computer at a centralized point. Control system engineers had been sketching out concepts of such distributed computer system composed of digital control and communication elements scattered about the plant since the middle of the 1960s^[10]. But, unfortunately, the technology to implement these concepts on a cost-effective manner was not available at that time.

Since the 1970s, as it is already familiar to all, the rapid development of the integrated circuit and the production in its ultimate form, the microprocessor or microcomputer, has changed the prospects of CDC, and indeed of all process control.

Besides their very small volume, there were three major and significant advantages of microprocessors and microcomputers over the earlier systems that greatly suited the needs of distributed control systems in the harsh industrial environment. These were high computational capability, high reliability and low cost. Together with the companion development of the necessary supporting techniques, such as building block design techniques, network data communication, standardized interfaces, color CRT display system, structured design of software and on-line diagnostics, etc., the microprocessors/microcomputers made the distributed system architecture practical, and since then the computer-based-digital control (CDC) has begun its all new and rapid development era.

The first practical and commercially successful distributed control system was developed by the Honeywell Company starting in 1969 to replace the earlier failed, centralized computer system and to solve the problem of DDC system reliability. This became the well-known TDC 2000 system and was widely followed by the other process control system vendors later in the 1970s and 1980s.

The new design included a set of small, widely distributed computer “boxes” containing one or more microprocessors. Each of these boxes controlled one or a very few loops. All of them were connected together by a single, very-high-speed data link permitting communications between each of the boxes and with a centralized console to

allow the operator to monitor the operation of each computer in the distributed system ^[11]. Figure 3. 5. 7 illustrates this concept. It should be compared directly with Figure 3. 5. 6 to give a dramatic appreciation of what was accomplished with this new concept (these sketches were adapted from Honeywell drawings).

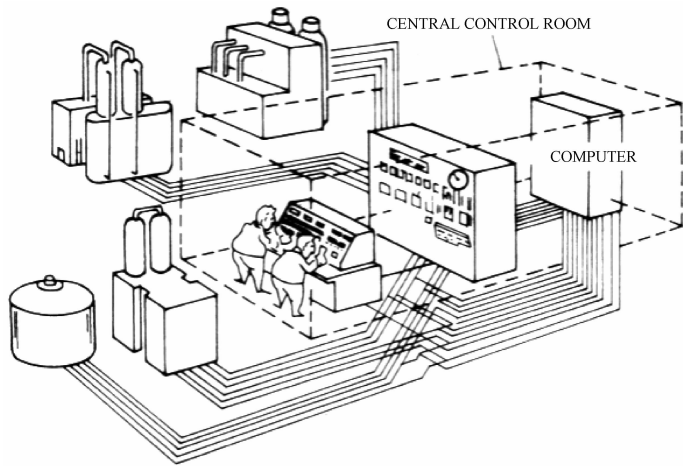


Figure 3. 5. 6 Sketch of a typical system with analog panel board and backup.

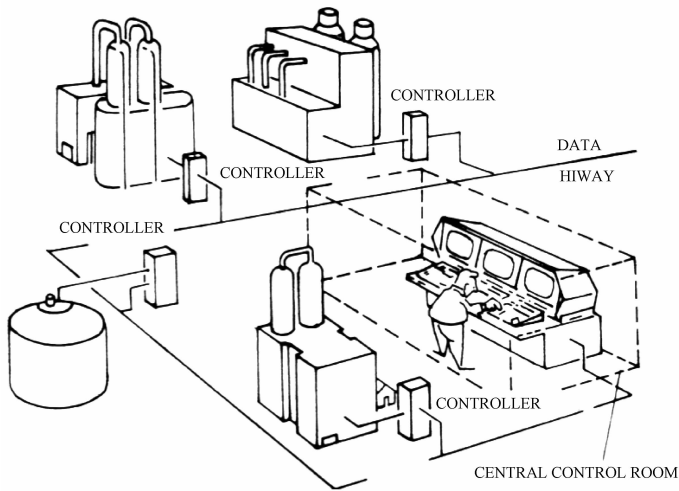


Figure 3. 5. 7 Concept of the microprocessor-based distributed direct digital computer control system (circa 1975 +).

3. 5. 2 Specialized English Words

single-loop controller	单(闭)环控制器,单控制器	signal read-out	信号读出(装置),信号示值(装置)
------------------------	---------------	-----------------	-------------------

regular pointer indication 常规指示
装置,常规显示装置
block diagram 方框图
field mounted 现场安装的,安装在现
场的
monitoring 监视
pneumatic 气动的
ratio instrument 比例式仪表
ammonia 氨
pointer 指针指示器,指针式仪表
strip chart 纸带记录仪
control set points 控制设定点,控制
设定值
refinery 炼油厂
supervisory control system 监控系统
direct digital control (DDC) 直接数
字控制
bypass 旁路,代替
specialized dedicated digital
controllers 专用数字控制器
ethylene 乙烯
computer box or mainframe 计算机
装置
central control-room 中央控制室
centralized computer system 集中式
计算机(控制)系统
architecture 结构
distributed computer systems 分布
式计算机系统
reliability 可靠性

microprocessor 微处理器
building block 积木,模块
color CRT 彩色显示器(CRT 为缩写
形式)
integrated circuit 集成电路
cathode ray tube 阴极射线管
process control 过程控制
on-line diagnostics 在线诊断
data-link 数据链路
plant 工厂,设备,(控制)对象
console 控制台,操作台
backup system 支持系统,后备系统
hot standby 热备用
input scanning 输入扫描
data base updating 数据库更新
control algorithm computation 控制
算法计算
real-time 实时
man-machine interfacing 人机接口
data logging 数据登录
computer-based control system 基于
计算机的控制系统
computer-based digital control (CDC)
基于计算机的数字控制
PID (Proportional, Integral and Deri-
vative controller mode) controller
PID(比例,积分,微分式)控制器,调
节器

3.5.3 Notes

[1] Note that in the Figure 3.5.1, Level 1A refers to those instruments which have no signal read-out readily available to the operator. such as some field-mounted controllers, ratio instruments. etc. 此句为省略了主语的祈使句。“Note that …”意为“值得注意的是…”,“that”引起宾语从句,“Level 1A”为从句的主语,而“instruments”为宾

语,由一个“which”引导的宾语从句修饰。全句可译为“值得注意的是,图中 1A 层指的是那些不具备向操作人员提供现成的信号读数的设备,比如某些安装在现场的控制器、比例式仪表等。”

[2] The first application of digital computers to industrial process was one developed for the plant monitoring of an electric power generation station at Sterlington, Louisiana, plant of Louisiana Power and Light Company in September, 1958, as shown in Figure 3.9.2. 句中的 one 为代词,指代“the first application of digital computers”。过去分词短语“developed for …”做定语,修饰“one”。全句可译为“首次用于工业过程的数字计算机是一个全厂监视系统,于 1958 年 9 月开发安装在路易斯安那电力与光源公司的一个位于路易斯安那州斯安特灵顿市的发电站,如图 3.5.2 所示。”

[3] “specialized dedicated digital controllers”意为专用数字控制器。“specialized dedicated”都是“专门的”、“专用的”的意思,这里是重叠修饰。

[4] While common today, these devices were not present in the very early DDC systems. 句中的“While common today”为“while these devices are common today”的省略形式,做状语。此句译为“尽管今天已十分普遍,但是最早的 DDC 系统中并没有这些设备。”

[5] The inherently severe performance requirements imposed upon earlier digital computers made it necessary for the manufacturers to develop much larger, much faster, and so, much more expensive computers to be built and used. 此句从结构上说是一个简单句,主语为“performance requirements, “the inherently severe”和“imposed upon the earlier digital computers”分别是“performance requirements”的前置定语和后置定语,“made”为谓语动词。“made it necessary”为典型的“动词+宾语+补语”结构。“it”指代实际宾语“to develop much larger … computers”,其中的“to be built and used”为“computers”的定语。全句可译为“对早期数字计算机所做的严格性能要求迫使制造商必须开发更快更大、因而也更贵的计算机以供制造和使用。”

[6] The biggest disadvantage of this architecture was that any failure of the computer itself could unavoidably shut down the entire system and the reliability of the computers was rather low at that time. 句中连词“and”相当于“while”,在此有转折之意,可译为“而”。全句可译为“这种结构的最大问题是,任何计算机自身的失效,都不可避免地会导致整个系统的关闭,而那时候计算机的可靠性又还相当低。”

[7] Another structural problem was that, because of the centralized computer location, the vast plant communication system required to bring the plant signals to the computer and return control signals to the field was also very expensive and was prone to the electrical noise problem. 这是一个系表结构的主从复合句,表语为从句形式。句中第一个“was”是主句的谓语动词,后两个“was”则是从句的两个谓语动词。“because of … location”是从句的状语。全句可译为“另一个结构性问题是,由于集中式计算机所处的位置,将全厂的信号传递给计算机,以及将控制信号反传给现场所需的庞大的通信系统也非常昂贵,还要受到电气噪声的干扰。”

[8] To compensate for the unreliability of the computer hardware, another “hot standby” computer or a complete analog backup system paralleling the DDC was needed, again greatly increasing its cost. 这是一个带有两个状语的句子，一个状语是句首的不定式短语“To compensate for the unreliability of the computer hardware”，表示目的；另一个状语是句末的现在分词短语“again greatly increasing its cost”，表示结果。“paralleling the DDC”是主语“a complete analog backup system”的后置定语。全句可译为“为了补偿计算机硬件的不可靠性，有必要另设一台‘热备用’计算机，或者设一整套与 DDC 并行的备用模拟系统，从而再一次大大增加了成本。”

[9] Again, to compensate these high costs, users and vendors alike attempted to squeeze as large as possible a project into the computer, further worsening the disadvantage of centralization and extremely, complicating its programming by requiring the computer to perform a wide variety of functions in real time: input scanning, database updating, control algorithm computation, data logging and man-machine interfacing, etc. 此句虽是一个简单句，但附加成分较复杂。试分析如下：“users and vendors”为主语，“attempted”为谓语动词。“to squeeze ... into ...”为“attempted”的不定式宾语。该句有三个状语：“again”和“to compensate these high costs”为不定式短语做状语，表示目的；“further worsening”和“complicating”为并列的现在分词短语做状语，表示结果。介词短语“by requiring...”进一步具体说明“worsening”和“complicating”的原因，而“as large as possible”用以修饰“a project”。全句可译为“为了弥补这些高昂的成本，用户和制造商都企图将一个尽可能大的方案塞进计算机中，这进一步凸现了集中控制方式的缺陷并使其编程过程极大地复杂化，因为这要求计算机必须在实时条件下执行大量的任务——输入扫描、数据库更新、控制算法运算、数据登录及人机接口等。”

[10] Control system engineers had been sketching out concepts of such distributed computer system composed of digital control and communication elements scattered about the plant since the middle of the 1960s. 注意句中三个过去分词的作用：“distributed”做“computer”的前置定语；“composed of ...”则是“computer”的后置定语；同样，“scattered about the plant”是“elements”的后置定语。全句可译为“自 20 世纪 60 年代中期以来，控制系统的工程师们就在勾勒像这样由分散在全厂的数控与通信部件组成的分布式计算机系统的概念。”

[11] All of them were connected together by a single, very-high-speed data link permitting communications between each of the boxes and with a centralized console to allow the operator to monitor the operation of each computer in the distributed system. 句中的介词短语“by a single very-high-speed ...”与不定式短语“to allow the operator ...”并列做状语。“with a centralized console”则是“to allow the operator ...”的状语。全句可译为“它们全都由一个单一的高速数据链路连接到一起，实现各计算机装置之间的通信，并通过一个集中控制台使操作员能监视分布式系统中每一台计算机的运行。”

3.5.4 Reference Translation

计算机控制系统的发展

工业自动化的历史是悠久的。由于生产的需要和技术的发展,1930 年代以来工业自动化已取得了重大进步。将计算机用于工业控制之前,标准的工业控制系统——其框图如图 3.5.1 所示——是由多个气动或电子单环控制器组成的。值得注意的是,图中 1A 层指的是那些不具备向操作人员提供现成的信号读数的设备,比如某些安装在现场的控制器、比例式仪表等。1B 层是能向操作人员提供规范指示的设备,例如纸带记录仪、指针式仪表等。这些就是后来的标准 PID(比例、积分和微分控制方式)调节器。

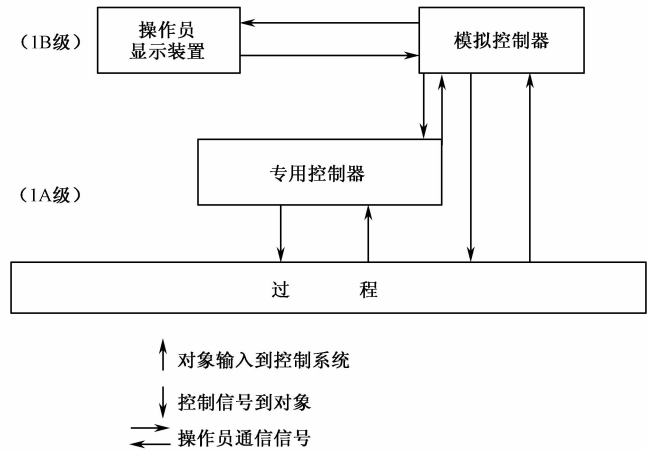


图 3.5.1 基本系统——模拟控制

首次用于工业过程的数字计算机是一个全厂监视系统,于 1958 年 9 月开发安装在路易斯安那电力与光源公司的一个位于路易斯安那州特灵顿市的发电站,如图 3.5.2 所示。接着在 1959 年 3 月和 1960 年 4 月,TEXACD 公司在得克萨斯的一家炼油厂——Monsanto 化学公司在路易斯安那的一家制氨厂,安装了首批计算机监视控制系统,其框图如图 3.5.3 所示。

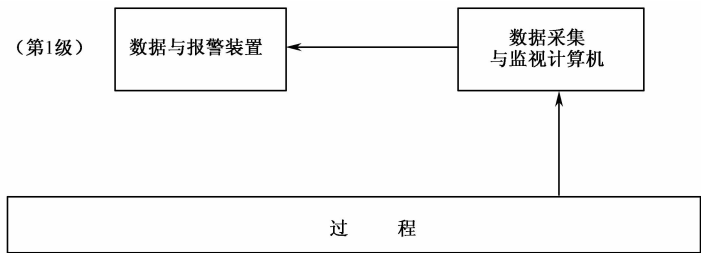


图 3.5.2 数据采集与监视系统

在这种计算机控制系统中,计算机计算控制设定值并把它们传给模拟控制器,并不直接控制生产过程。模拟控制器仍然是过程控制的主要方式。

计算机控制系统下一步的发展必然是用计算机取代图 3.5.3 中的模拟控制器,直接对对象进行控制。这就是所谓的直接数字控制(DDC)。它的首次出现是由 Monsanto 公司于 1962 年 3 月在得克萨斯市的乙烯厂完成的,如图 3.5.4 所示。请注意,图 3.5.4 包括了可能出现的专用数字控制器。尽管今天已十分普遍,但是最早的 DDC 系统中并没有这些设备。当时所有的工作都是由 1B 级的数字计算机来完成的。

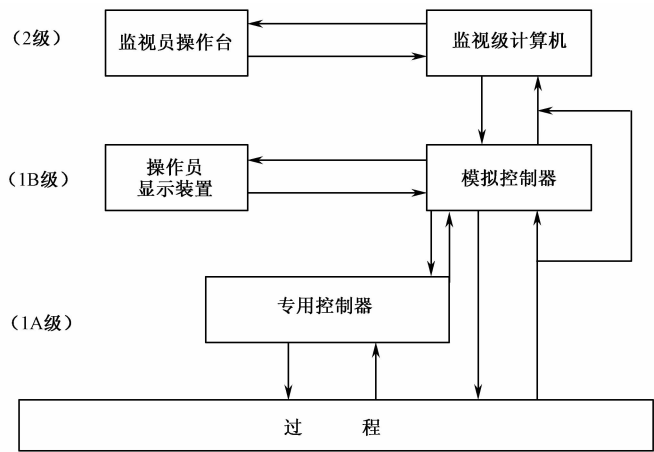


图 3.5.3 仅用做监视的辅助性数字控制

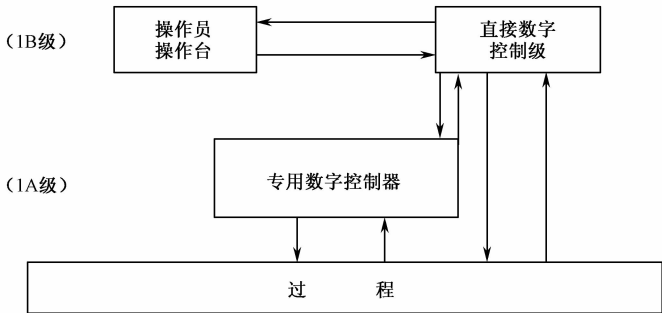


图 3.5.4 仅用做基本数字控制

人们应知道,计算机自身最初的发展起始于 1946 年,比计算机用于工业控制系统仅早了 12 年。尽管计算机的工程应用的强大能力和潜力早已被广泛认可;而且,在 20 世纪 50 年代和 60 年代,计算机科学技术的发展也很迅速,但当时(20 世纪 60 年代中期)它仍处在发展的早期阶段。对早期数字计算机所做的严格性能要求迫使制造商必须开发更快更大、因而也更贵的计算机以供制造和使用。为了有助于平衡这种高成本,制造商极力主张将所有的计算机功能,包括监控和 DDC 两者,全都集中到位于中央控制室的一台计算机之中,如图 3.5.5 和图 3.5.6 所示。我们把这称为集中式计算机控制系统。

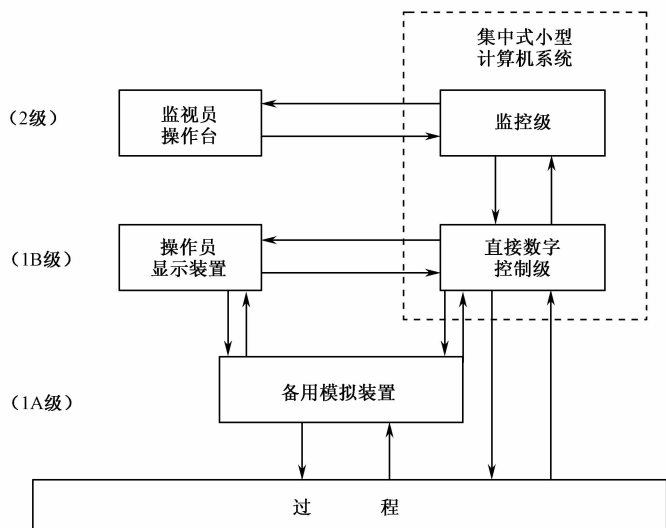


图 3.5.5 完全的辅助数字控制,带备用模拟控制的监控加直接数字控制由集中式小型计算机进行控制运算

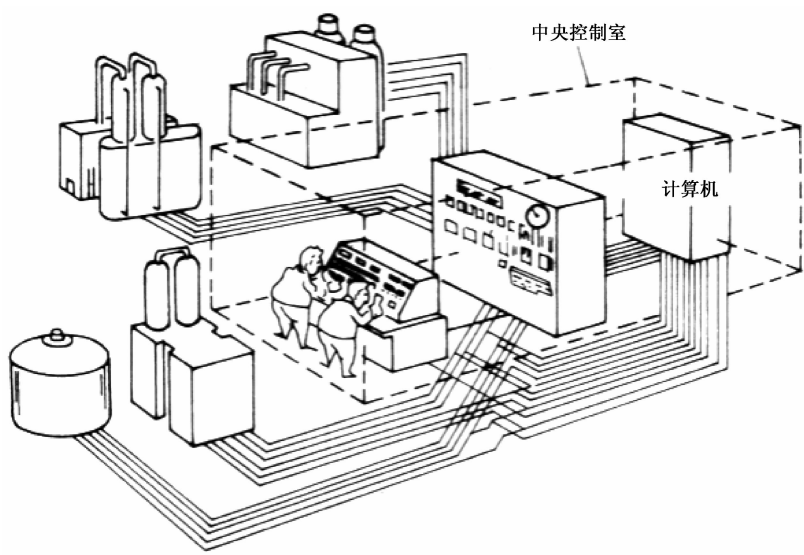


图 3.5.6 配备模拟台及备用装置的典型系统示意图

尽管集中式计算机控制系统比早先的控制系统有很大的优点,但也受到不少缺点的困扰。这种结构最大的问题是,任何计算机自身的失效,都不可避免地会导致整个系统的关闭,而那时候计算机的可靠性又还相当低。另一个结构性问题是,由于集中式计算机所处的位置,将全厂的信号传递给计算机,以及将控制信号反传给现场所需的庞大通信系统

也非常昂贵,而且还要受到电气噪声的干扰。

为了补偿计算机硬件的不可靠性,有必要另设一台“热备用”计算机,或者设一整套与 DDC 并行的备用模拟系统,从而再一次大大增加了成本。为了弥补这些高昂的成本,用户和制造商都企图将一个尽可能大的方案塞进计算机中,这进一步凸现了集中控制方式的缺陷并使其编程过程极大地复杂化,因为这要求计算机必须在实时条件下执行大量的任务——输入扫描、数据库更新、控制算法运算、数据登录及人机接口等。

这些问题导致工厂管理层对计算机失去了热情,并使得计算机控制系统的安装完全停顿不下来。这种局面一直持续到 20 世纪 70 年代中期。

情况已经变得很清楚,需要有一种新的方法,这种方法在速度、可靠性和存储容量等方面具备确保提供满足用户需求和期望的能力。也许需要一种新的结构——计算机必须分散到工厂的各处,以避免将一台计算机放在一个集中点所带来的局限。自 20 世纪 60 年代中期以来,控制系统工程师们就在勾勒像这样的由分散在全厂的数控与通信部件组成的分布式计算机系统的概念。但是很遗憾,那个时候还不具备能以一种在经济上合理的方式实现这些理念的技术。

20 世纪 70 年代以来,人所尽知,集成电路及其终极形式——微处理器和微型计算机——生产的迅速发展,改变了 CDC(计算机数字控制),说实在的,是整个过程控制的前景。除了体积极小之外,微处理器和微型计算机比以前的计算机具有三个主要的大优点,在严酷的工业环境下非常适合分散式控制系统的需要。这些优点是高计算能力、高可靠性和低成本。和一些必备支持技术如模块化设计技术、网络数字通信、标准化接口、彩色 CRT 显示系统、软件结构化设计以及在线诊断等的共同发展的配合下,微处理器/微型计算机使得分布式系统结构实用化了。从此,CDC 开始了它全新的快速发展时代。

第一个实用而且在商业上成功的分布式控制系统是由 Honeywell 公司于 1969 年开始开发的,意在取代早期不成功的集中式计算机系统并解决 DDC 系统的可靠性问题。这就是著名的 TDC 2000 系统。后来在 20 世纪 70 年代和 20 世纪 80 年代,其他过程控制系统制造商都纷纷跟进。

这个新设计包含一组分得很散的小计算机“匣子”,匣内有一个或多个微处理器。每个匣子控制一个或很少几个控制闭环。它们全都由一个单一的高速数据链路连接到一起,实现各计算机装置之间的通信,并通过一个集中控制台使操作员能监视分布式系统中每一台计算机的运行。

图 3.5.7 描述了这一概念。把它和图 3.5.6 直接对比,应能对这个新概念所能实现的东西获得形象的首肯(这两幅图源自 Honeywell)。

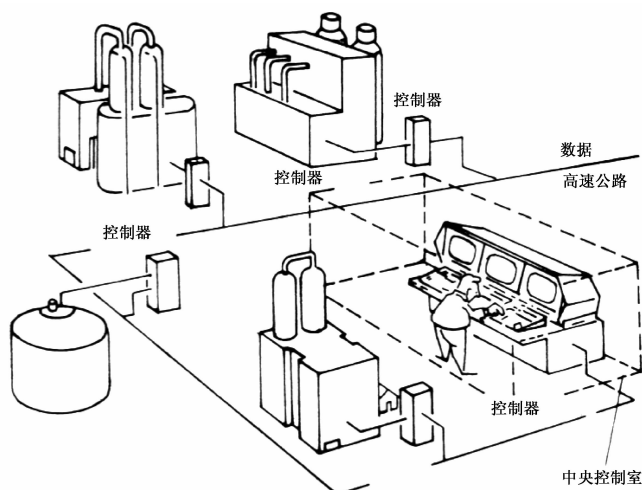


图 3.5.7 以微处理器为基础的分布式直接数字计算机控制系统示意图(1975 年前后)

3.5.5 Reading Materials

Computer Control

Among the many spectacular (引人注目的) aspects of the growth of the digital computer field (现场) has been its application to industrial process control. From very modest beginnings in 1959 and 1960, the field has now progressed until it is an indispensable part of many new industrial plants, particularly in the steel industry, petrochemical plant operations, refineries, and several others. Due to small memory sizes, relatively slow machine speeds, and poor reliability, early applications were in a data-logging or operator's guide configuration. Increased knowledge of process requirements led to the closed loop supervisory or steady-state, economic optimization phase (时期, 阶段) of computer control. However, because of limitations in plant excursion ranges (范围), results here were small, although positive. Dynamic control by digital computer has come only recently with direct digital control. This was made possible by the emergence of a third-generation group of small, fast, highly-reliable, and relatively inexpensive machines. These machines have shown themselves to be economically attractive media for plant control, with a resulting major increase in applications in the digital computer control field. Overall plant or hierarchy control, where several levels of computers operate together to control a large plant complex or a whole company, is now under active study by many groups and preliminary (初步的, 尝试性的) installations of this type are already on order (预定) or being installed. This area promises to grow rapidly in the next few years, along with the other phases of computer control. This paper reviews the areas of digital plant control, pointing out

areas of difficulty and of successes in each. It lists some requirements for future progress in terms of both equipment and software requirements. It also attempts to predict some probable future areas and rates(速度) of development.

3.6 General Concepts of Hierarchical Control

3.6.1 Text

As we can see from Figure 3. 6. 1, three or more levels of control devices, each with distinct duties, form a computer system hierarchy, where the upper-level computers depend on lower level devices for process data, and the lower level systems in turn depend on the higher level systems for ever more sophisticated control functions, such as overall plant optimization^[1]. This architecture then includes production scheduling and management information functions as well as the process control functions, forming a total plant control system. Figure 3. 6. 2 shows the control structure for a continuous processing industry such as a petroleum refinery or a chemical plant. Figure 3. 6. 3 shows the structure for a discrete manufacturing plant. These two diagrams are exactly functional equivalent except for the use of different nomenclatures for similar functions by each type of industry.

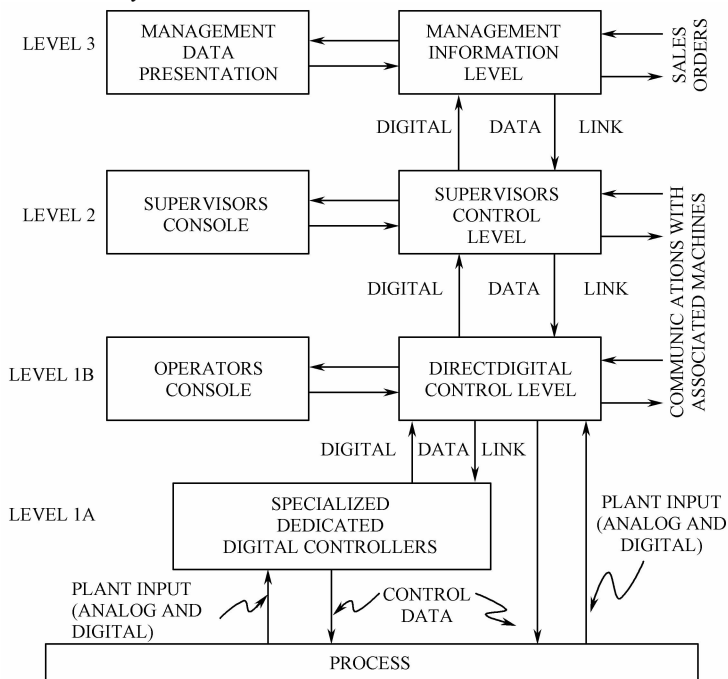


Figure 3. 6. 1 Outline of hierarchy-organization for a complete process computer control system.

With the steady development of CDC, it has been very clear that “a hierarchically organized distributed and computational capability must be the logical structure of the control system for any plant, regardless of industry.”The concept of the hierarchy is of great importance to computer integrated manufacturing (CIM).

In the broadest view of industrial automation an overall automatic control system for any manufacturing plant regardless of the industry should have the following capabilities:

1. An effective dynamic control of each operating unit of the plant to assure that it is operating at its maximum efficiency of production capability, product quality and/or of energy and materials utilization based on the production level set by the scheduling and supervisory functions listed below^[2]. This thus becomes the control enforcement component of the system. This control reacts directly to compensate for any emergencies which may occur in its own unit.

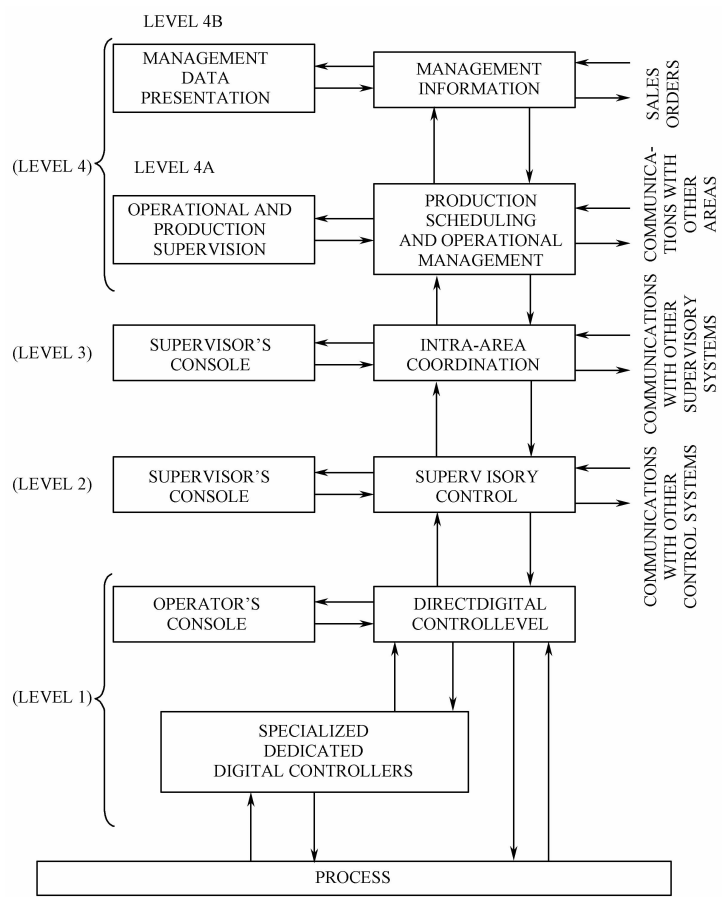


Figure 3. 6. 2 Assumed hierarchical computer control structure for an industrial plant (continuous process such as petrochemicals).

2. A supervisory and coordinating system which determines and sets the local production level of all units working together between inventory locations in order to continually improve (i. e, optimize) their operation^[3]. This system assure no unit is exceeding the general area level of production and thus using excess raw materials or energy. This system also responds to the existence of emergencies or upsets in any of the units under its control in cooperation with those units' dynamic control systems to shut down or systematically reduce the output in these and related units as necessary to compensate for the emergency ^[4]. In addition, this system is responsible for the efficient reduction of plant operational data from the dynamic control units, described just above, to assure its availability for use by any plant entity requiring access to it as well as its use for the historical data base of the plant ^[5].

3. An over-all production control system capable of carrying out the scheduling functions for the plant from customer orders or management decisions so as to produce the required products for these orders at the best (near optimum) combination of customer service and the use of time, energy, inventory, manpower and raw materials suitably expressed as cost functions^[6].

4. A method of assuring the overall reliability and availability of the total control system through fault detection, fault tolerance, redundancy, uninterruptible power supplies, maintenance planning and other applicable techniques built into the system's specification and operation^[7].

Because of the ever-widening scope of authority of each of the fast three requirements in turn, they effectively become the distinct and separate levels of a superimposed control structure, one on the top of the other^[8]. Thus, a multi-level hierarchical architecture is readily evolved, as we have seen, for example, in Figure 3. 6. 2 where we can easily match each level with these three requirements just mentioned ^[9]:

1. The first capability is carried out in Level 1A and 1B.
2. The second capability is carried out in Level 2.
3. The third capability is carved out in Levels 3 and 4.

In fact, the organizational structure of industrial production and of the management of a factory is itself naturally divided into levels according to their own different functions and their relations—from the low level process running to the high level management. So, it is a logical approach to use a hierarchical control system to meet the needs of a hierarchical physical system.

Also the hierarchy is a very effective concept of vertical decentralization or distribution here applied to the idea of distributed control. Here we will lay stress on one aspect of it as follows:

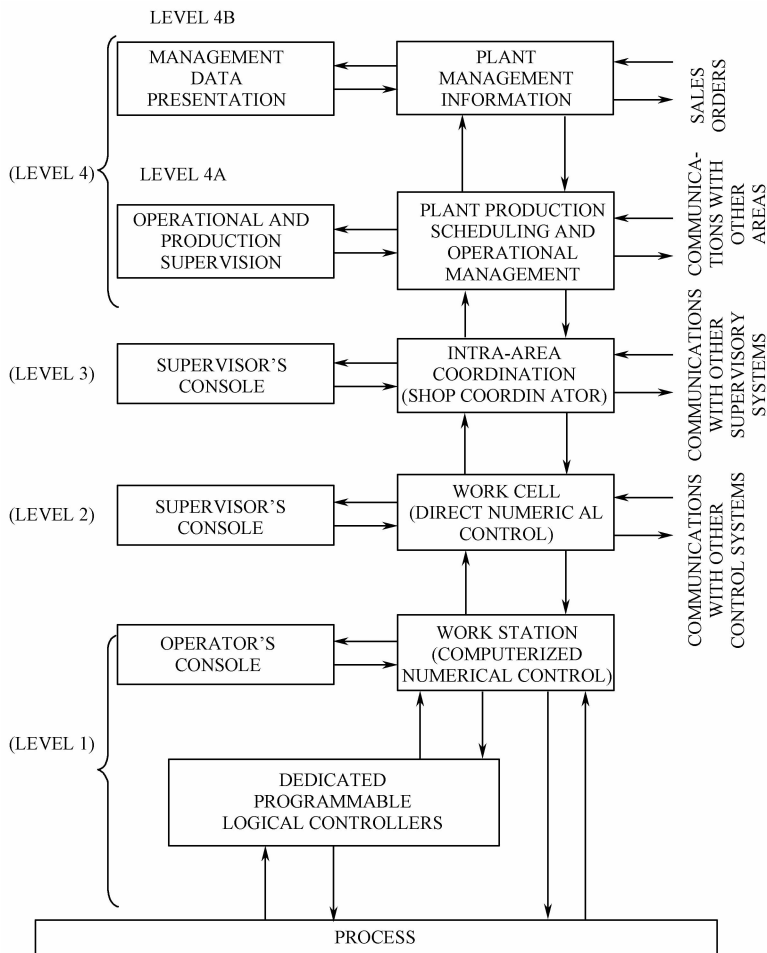


Figure 3.6.3 Assumed physical hierarchy computer system structure for a large manufacturing complex (computer integrated manufacturing system (CIMS)).

Since the total system is decomposed into different levels of smaller and more manageable subsystems (normally having the classical type of feedback control structure on the lowest level, while the higher levels involve more sophisticated functions), it is more flexible and easier to develop, install and maintain than any earlier system with the same capability. The same applies when expanding the whole system^[10]. For instance, a control system with the lower two or three levels can be set up at first. When enough experience has been obtained, the more complicated higher levels can be directly added on top of them without up-setting the existing system.

This is of great significance when we note that, because of the complexity, very

few such systems containing all the six levels as in Figure 3. 6. 4, which is a synthesis of Figure 3. 6. 2 and Figure 3. 6. 3, have yet been achieved at the present time. So the hierarchical architecture gives the system designers a really effective, flexible and convenient way to reach their ultimate dream step by step.

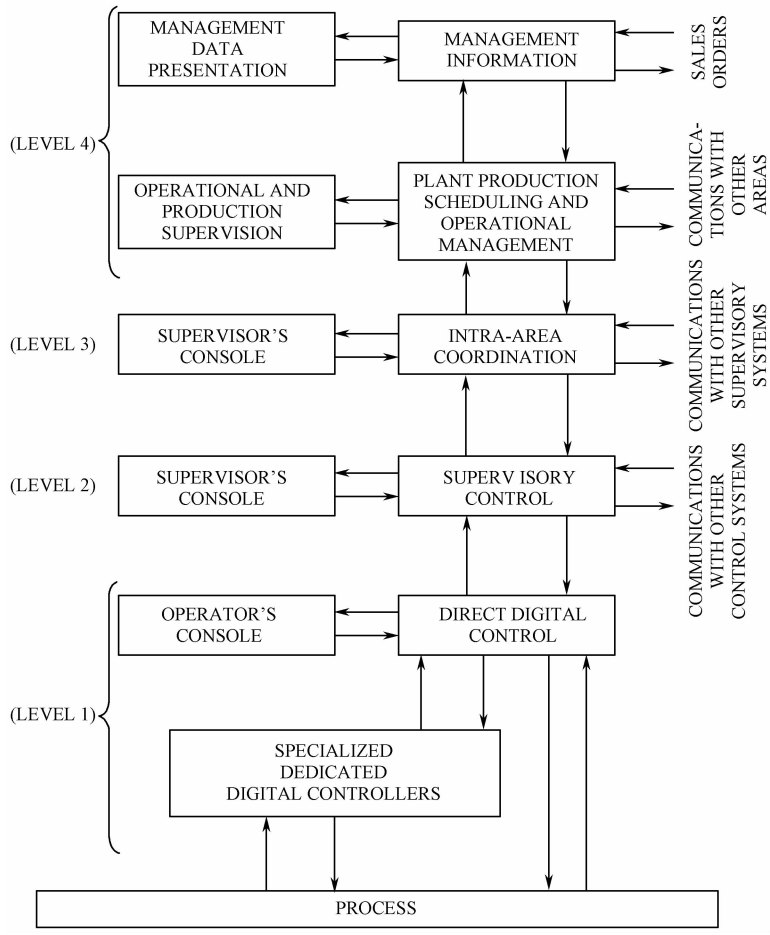


Figure 3. 6. 4 Definition of the real tasks of the hierarchical computer control system.

3. 6. 2 Specialized English Words

- | | | | |
|---------------------|------|----------------------------|------|
| hierachical | 分级控制 | inventory locations | 设备线 |
| sales | 销售 | hierarchy | 分级结构 |
| orders | 订货 | overall plant optimization | 全厂优化 |
| control enforcement | 控制系统 | production scheduling | 生产计划 |
| component | 组成部分 | management information | 管理信息 |

total plant control system 全厂控制系统
 continuous process 连续(生产)过程
 intra-area coordination 区域内协调
 operational management 运行管理
 specification 技术规范,技术特点
 workstation 工段,车间,工作站
 digital numerical control 直接数字控制
 operational data 运行数据
 customer order 用户订单
 historical data base 历史数据库,历史资料库
 cost functions 成本函数
 fault detection 差错
 fault tolerance 容错
 redundancy 冗余
 uninterruptable power supplies

(UPS)不间断电源
 authority 职责
 superimposed 重叠的,叠加的
 discrete 离散的
 computational 计算的
 nomenclatures 名词,术语
 computer integrated manufacturing (CIM)计算机集成制造(系统)
 production scheduling and management information 生产计划与管理信息
 control computation and control enforcement 控制计算与控制实施
 programmable logic controllers 可编程逻辑控制器
 vertical decentralization 垂直(方向)的分散和分布

3.6.3 Notes

[1]As we can see from Figure 3.6.1, three or more levels of control devices, each with distinct duties, form a computer system hierarchy, where the upper-level computers depend on lower level devices for process data, and the lower level systems in turn depend on the higher level systems for ever more sophisticated control functions, such as overall plant optimization. 此句结构比较复杂,从句“as”到谓语动词“form”之间的插入部分“each with distinct duties”为主语的定语;句子的后半部分是由“where”引导的定语从句,修饰主语中的宾语“hierarchy”。全句可译为“从图 3.6.1 我们可以看到,三层或更多层的控制设备——每层都有明确的职能——构成了一个计算机系统层级结构。在这个结构中,上层的计算机通过下层的设备获得过程的数据;反之,下层的计算机则通过上层的计算机实现更精深的控制,例如全厂优化控制。”

[2]An effective dynamic control of each operating unit of the plant to assure that it is operating at its maximum efficiency of production capability, product quality and/or of energy and materials utilization based on the production level set by the scheduling and supervisory functions listed below. 句中,“an effective dynamic control of each operating unit of the plant”为名词性短语,后跟不定式短语“to assure …”做它的定语,而“assure”又带一个由“that”引导的宾语从句。注意从句中有三个过去分词短语:“based on the production level”为“is operating”的状态语;“set by the scheduling and supervisory functions”为“production level”的后置定语,而“listed below”则是“the scheduling and

supervisory function”的后置定语。这段文字可译为“工厂的每一个运行单元都有一个有效的动态控制系统,该系统能确保在达到一定生产水平的前提下,使生产产量和产品质量、能源及材料消耗等都处在最高效率状态之下,而生产水平由下面将谈到的计划与监控模块所设定。”

[3] A supervisory and coordinating system which determines and sets the local production level of all units working together between inventory locations in order to continually improve (i. e, optimize) their operation. 这一小段与注[2]相似,也是名词性短语带一个后置定语,不同的是后置定语为“which”引导的从句,从句中的现在分词短语“working ... locations”是“all units”的定语,而“in order to ...”则是“working”的状语。全句可译为“有一监控协调系统,负责确定及设定所有单元的局部生产水平,使位于生产线上的各单元共同协调,以便连续改进(即优化)其运行状态。系统应确保各单元不得超越该单元的生产水平从而耗用过多的原材料和能源。”

[4] This system also responds to the existence of emergencies or upsets in any of the units under its control in cooperation with those units' dynamic control systems to shut down or systematically reduce the output in these and related units as necessary to compensate for the emergency. 这是一个结构为“主(This system)+谓(responds)+宾(to the existence of ...under its control)”的简单句,但附加成分较复杂。“in cooperation with those units' dynamic control systems”为方式状语,紧跟其后的“to shut down or ... reduce”则是并列的不定短语,做目的状语,而“to shut down or systematically reduce”又有自己的状语“as necessary to compensate for the emergency”。全句可译为“该系统通过与相应单元的动态控制系统的协调,对它所控制的任何单元中出现的紧急情况做出反应,以便根据需要关闭或系统地减少这些单元的输出,来应对紧急情况。”

[5] In addition, this system is responsible for the efficient reduction of plant operational data from the dynamic control units, described just above, to assure its availability for use by any plant entity requiring access to it as well as its use for the historical data base of the plant. 这句话的结构与上句相似。要注意现在分词短语“requiring access to it”是“plant entity”的后置定语。句中的两个“its”均为“the efficient reduction of plant operational data”的所有格。全句可译为“此外,该系统还负责对来自上文刚谈到的动态控制单元的运行数据进行有效压缩,以满足工厂中任何部门访问该数据的需要,并将这些数据用于建设工厂的历史资料库。”

[6] An over-all production control system capable of carrying out the scheduling functions for the plant from customer orders or management decisions so as to produce the required products for these orders at the best (near optimum) combination of customer service and the use of time, energy, inventory, manpower and raw materials suitably expressed as cost functions. 这段文字的结构很简单,可以认为“capable of ...”之前省略了“which is”,所以“capable of”以下均为“an over-all production control system”的定语,其中的“so as to...”为“carrying out”的状语,句末的“expressed as cost

functions”为“combination”的定语。全句可译为“有一个总体的生产控制系统,能根据客户订单或管理决策实施工厂计划功能,把为顾客的服务和所耗用的时间、能源、设备、人力及原材料这两者结合起来,以最佳方式(近似最优)生产所需的产品。这里所用的时间、能源、设备、人力及原材料是和成本相关联的。”

[7] A method of assuring the overall reliability and availability of the total control system through fault detection, fault tolerance, redundancy, uninterruptible power supplies, maintenance planning and other applicable techniques built into the system's specification and operation. 句末的“built into ...”为“other applicable techniques”的定语。全句可译为“能通过查错、容错、冗余、不间断电源供电、维护策略及纳入系统技术规范及系统运行的其他有效技术手段,确保整个控制系统的可靠性和可用性。”

[8] Because of the ever-widening scope of authority of each of the fast three requirements in turn, they effectively become the distinct and separate levels of a superimposed control structure, one on the top of the other. 此句为简单句。句末的“one on the top of the other”可视为插入成分,对前面谈到的情况做进一步的补充解释。全句可译为“由于前三项要求所对应的职责范围依次不断扩展,它们必然演变为一个控制体系中的明确而分开的层级,一级叠加在另一级之上。”

[9] Thus, a multi-level hierarchical architecture is readily evolved, as we have seen, for example, in Figure 3. 6. 2 where we can easily match each level with these three requirements just mentioned. 这是一个复合句。从“as we have seen”至句末为状语从句,从句中的介词短语“in Figure 3. 6. 2”为状语,而“where ...”又为介词宾语“Figure 3. 6. 2”的定语从句。全句可译为“这样,就自然地形成了一个如我们已在图 3. 6. 2中所看到的多层分级结构,从中可以很容易地将各层与刚谈到的三项要求一一对应起来。”

[10] The same applies when expanding the whole system. 句中,“the same”为主语,“applies”为谓语,意为“适用”,“when”引起状语从句,“expanding”前省略了“we are”。全句可译为“当整个系统扩展时情况也一样。”

3. 6. 4 Reference Translation

分级控制的一般概念

从图 3. 6. 1 我们可以看到,三层或更多层的控制设备——每层都有明确的职能——构成了一个计算机系统层级结构。在这个结构中,上层的计算机通过下层的设备获得过程的数据;反之,下层的计算机则通过上层的计算机实现更精深的控制,例如全厂优化控制。于是,这个结构体系包括了生产计划和管理信息功能以及过程控制功能,组成了一个全厂的控制系统。图 3. 6. 2 显示的是像石油炼油厂或化工厂这样的连续过程行业的控制结构。图 3. 6. 3 显示的则是离散制造厂的控制结构。这两幅图除了因行业不同而用了不同的名词术语外,在功能上是完全一样的。

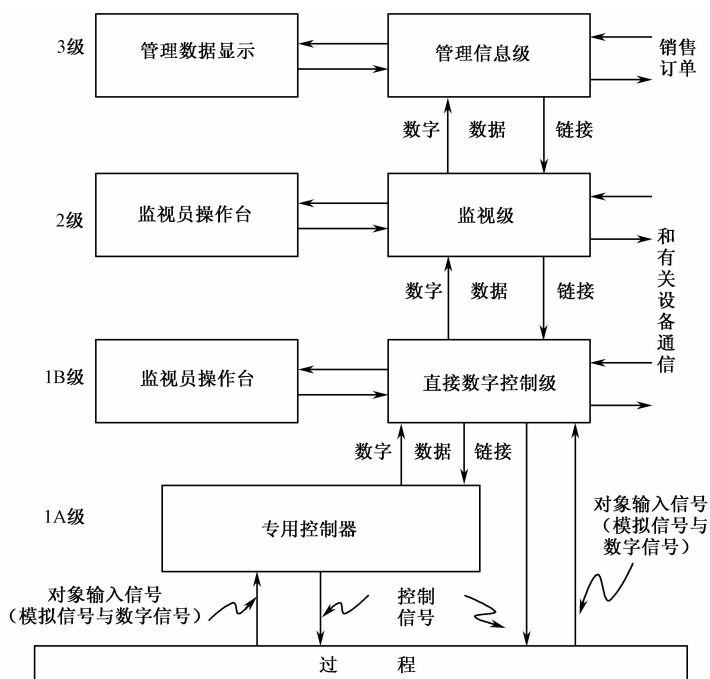


图 3.6.1 完整的计算机过程控制系统分级示意图

随着 CDC 的稳步发展,已经很清楚,“一个分级组织的、分布式的计算机系统是任何工厂的控制系统必然的结构,与工业行业无关”。这一分级的概念对于计算机集成制造 (CIM)具有重大意义。

从最广义的工业自动化观点看,一个与行业无关、适合任何制造工厂的全厂自动控制系统应当具备以下功能:

1. 工厂的每一个运行单元都有一个有效的动态控制系统,该系统能确保在达到一定生产水平的前提下,使生产产量和产品质量、能源及材料消耗等都处在最高效率状态之下,而生产水平由下面将谈到的计划与监控模块所设定。这就形成了系统的控制实施部分。这一部分将对本单元内可能发生的任何紧急情况直接做出反应,进行调整补偿。
2. 有一监控协调系统,负责确定及设定所有单元的局部生产水平,使位于生产线上的各单元共同协调,以便连续改进(即优化)其运行状态。系统应确保各单元不得超越该单元的生产水平从而耗用过度的原材料和能源。该系统通过与相应单元的动态控制系统的协调,对它所控制的任何单元中出现的紧急情况做出反应,以便根据需要关闭或系统地减少这些单元的输出,来应对紧急情况。此外,该系统还负责对来自上文刚谈到的动态控制单元的运行数据进行有效压缩,以满足工厂中任何部门对访问该数据的需要,并将这些数据用于建设工厂的历史资料库。
3. 有一个总体的生产控制系统,能根据客户订单或管理决策实施工厂计划功能,把为

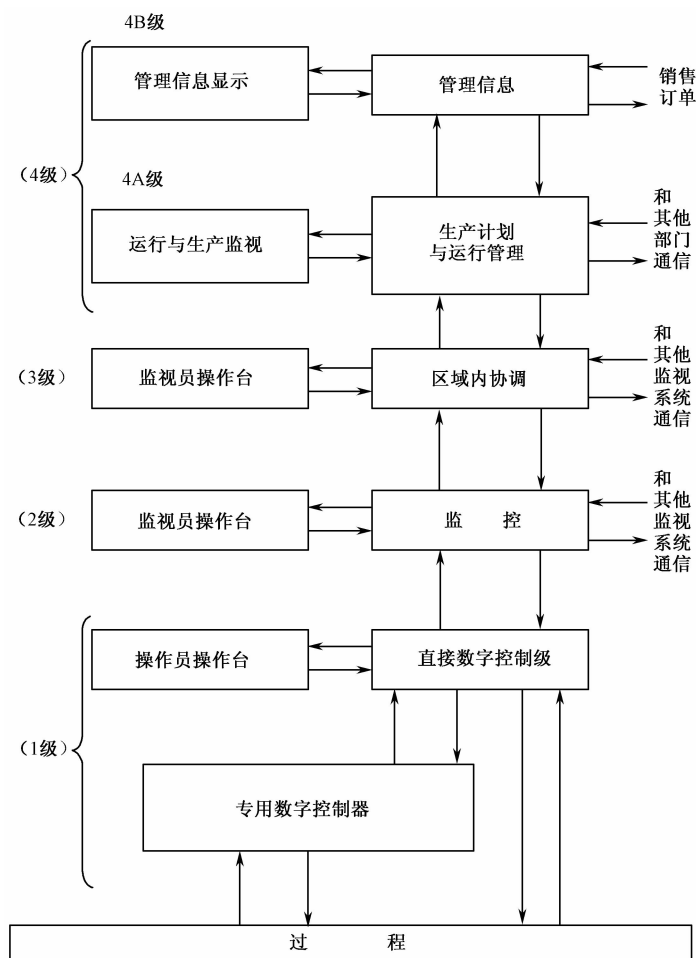


图 3.6.2 工厂分级计算机控制结构示意图(连续过程, 如石油化工)

顾客的服务和所耗用的时间、能源、设备、人力及原材料这两者结合起来,以最佳方式(近似最优)生产所需的产品。这里所用的时间、能源、设备、人力及原材料都是和成本相关联的。

4. 能通过查错、容错、冗余、不间断电源供电、维护策略及纳入系统技术规范及系统运行的其他有效技术手段,确保整个控制系统的可靠性和可用性。由于前三项要求所对应的职责范围依次不断扩展,它们必然演变为一个控制体系中的明确而分开的层级,一级叠加在另一级之上。这样,就自然地形成了一个如我们已在图 3.6.2 中所看到的多层分级结构,从中可以很容易地将各层与刚谈到的三项要求一一对应起来。

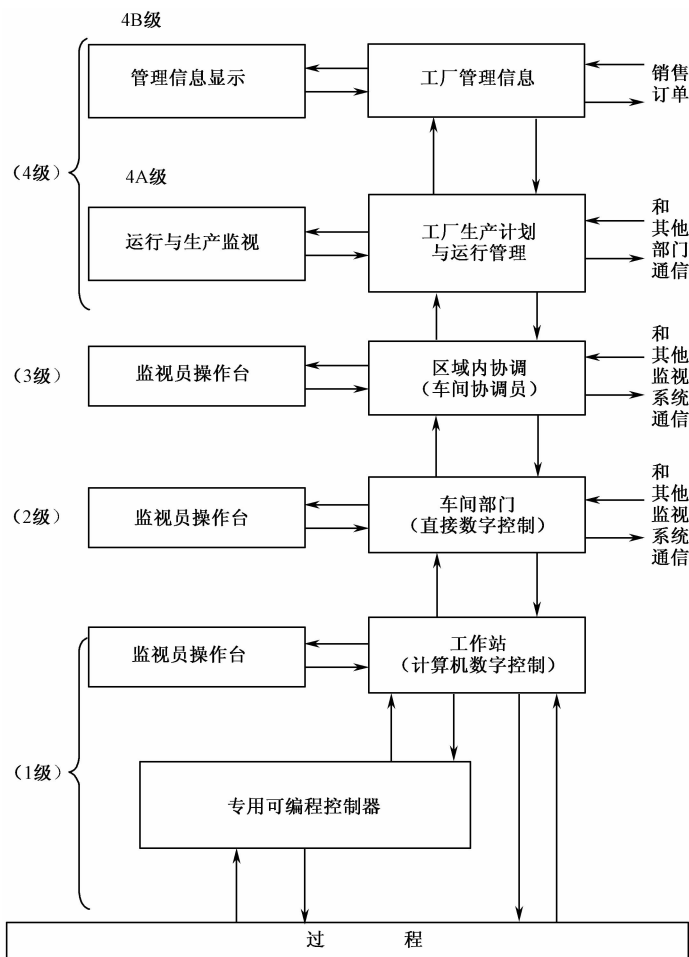


图 3.6.3 大型复杂制造系统分级计算机系统示意[计算机集成制造系统(CIMS)]

1. 第一个功能由 1A 级和 1B 级完成。
2. 第二个功能由 2 级完成。
3. 第三个功能由 3 级和 4 级完成。

实际上,一个工厂的工业生产和管理的组织结构本身就是根据自身不同的功能和相互关系而自然分级的——从下层的过程控制一直到上层的管理。所以,用一个分级的控制系统来满足一个分级的物理系统的需要是合乎逻辑的。

对于此间应用分布式控制理念的垂直方向上的分散,或者说分布而言,分级思想也是非常有用的。这里仅就其中一个方面强调如下。

整个系统可以通过分解为不同的层级而得到更小、更便于掌握的子系统(通常典型结构是反馈控制系统在最下层,而较高的层次实施更复杂的功能)。这样,和以往任何具有

同样功能的系统相比,开发、安装和维护起来都更为灵活和容易。当整个系统扩展时情况也一样。例如,可以首先建立包括较低的二层或三层的控制系统,待积累了足够的经验之后,功能更复杂的高层级可以直接加之其上而不会对已有系统造成不良影响。

由于事情的复杂性,目前图 3. 6. 4——它是图 3. 6. 2 和图 3. 6. 3 的综合——所示的包括全部 6 个层级的系统还建立得很少,注意到这一情况我们就能知道上述观点的重大意义。所以说,分级结构给系统设计师们开辟了一条真正灵活有效的方便之路,一步一步地去实现他们的最终梦想。

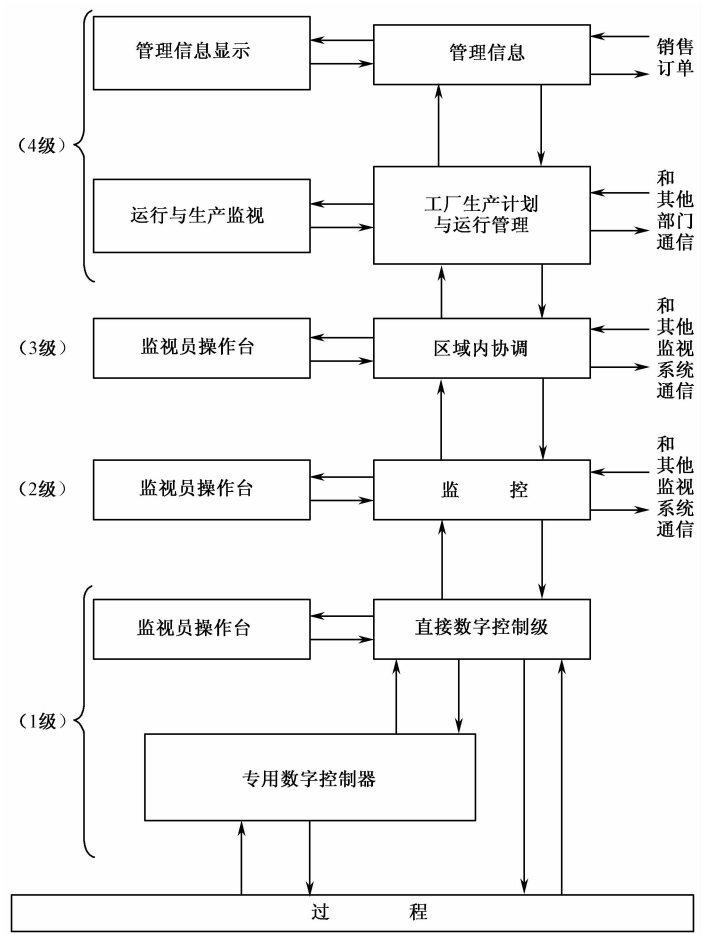


图 3. 6. 4 大型复杂制造系统分级计算机系统示意[计算机集成制造系统(CIMS)]

3. 6. 5 Reading Materials

The Industrial Robot

George Charles Devol is often called the father of robotics. He invented the first

industrial robot, the Unimate, in 1954. A few years later, Devol and Joseph F. Engelberger formed the first robot company, Unimation. In 1960, Unimation was purchased by Condec Corporation. General Motors installed the Unimate for die casting handling and spot welding in 1961.

Modern industrial robot arms continued to evolve in the 1960's and 70's. In 1963, the six-jointed(有 6 个关节的) Rancho Arm was created to assist handicapped(残疾人). This was followed by the tentacle arm(触角机械臂), designed by Marvin Minsky in 1968. It was able to lift a person and had 12 joints.

It was the 1969 Stanford Arm that eventually led to commercial arm production. The Stanford Arm was one of the first electronically powered, computer-controlled arms. By 1974, it reached a level of sophistication(精巧的程度) where it could assemble a Model T water pump.

The Stanford Arm was followed by the Silver Arm in 1974. The Silver Arm was created by MIT's David Silver to perform precise assembly using touch and pressure sensors and a microcomputer. These arms led to Victor Scheinman, the inventor of the Stanford Arm, to form Vicarm, Inc. in 1974 to manufacture industrial robotic arms. Scheinman was instrumental in the creation of the PUMA (programmable universal manipulator for assembly) for Unimation. In 1977, the European robot company ASEA, built two sizes of industrial robots.

Influences

The development of the computer directly influenced the advancement of industrial robotics. The automotive industry was another contributor. In the 1980's, automotive companies showered robotic companies with investments(纷纷向机器人公司投资).

But the enthusiasm and funding were not always matched with understanding. General Motors Corporation spent more than 40 billion on new technology in the 1980's. But a lack of understanding led to costly robot fiascos(惨败). In 1988, robots at the Hamtramck Michigan plant wreaked(报复) havoc-smashing windows(砸玻璃窗) and painting(涂漆) one another.

The premature introduction of robotics created financial instability. The robotics industry has only recently regained(重新得到) mid-1980 revenue(收入, 经费) levels. The American robotics market disappeared as Japanese and European bought up companies.

Part 4 Automatic Control Theory

4.1 History of Automatic Control

4.1.1 Text

The use of feedback to control a system has a fascinating history. The first applications of feedback control appeared in the development of float regulator mechanisms in Greece in the period 300 to 1 B. C.. The water clock of Ktesibios used a float regulator. An oil lamp devised by Philon in approximately 250 B. C. used a float regulator in an oil lamp for maintaining a constant level of fuel oil. Heron of Alexandria, who lived in the first century A. D. , published a book entitled *Pneumatica*, which outlined several forms of water-level mechanisms using float regulators^[1].

The first feedback system to be invented in modern Europe was the temperature regulator of Cornelis Drebbel (1572-1633) of Holland. Dennis Papin (1647-1712) invented the first pressure regulator for steam boilers in 1681. Papin's pressure regulator was a form of safety regulator similar to a pressure-cooker valve.

The first automatic feedback controller used in an industrial process is generally agreed to be James Watt's fly ball governor, developed in 1769 for controlling the speed of a steam engine. The all-mechanical device, shown in Figure 4. 1. 1, measured the speed of the output shaft and utilized the movement of the fly ball with speed to control the valve and therefore the amount of steam entering the engine. As the speed increases, the ball weights rise and move away from the shaft axis, thus closing the valve. The flyweights require power from the engine to turn and therefore cause the speed measurement to be less accurate.

The first historical feedback system, claimed by Russia, is the water-level float regulator said to have been invented by I. Polzunov in 1765. The level regulator system is shown in Figure 4. 1. 2. The float detects the water level and controls the valve that covers the water inlet in the boiler.

The period preceding 1868 was characterized by the development of automatic control systems through intuition and invention. Efforts to increase the accuracy of the control systems led to slower attenuation of the transient oscillations and even to unstable systems. It then became imperative to develop a theory of automatic control. J. C. Maxwell formulated a mathematical theory related to control theory using a differential equating model of a governor^[2]. Maxwell's study was concerned with the

effect various system parameters had on the system performance^[3]. During the same period, I. A. Vyshnegradskii formulated a mathematical theory of regulators.

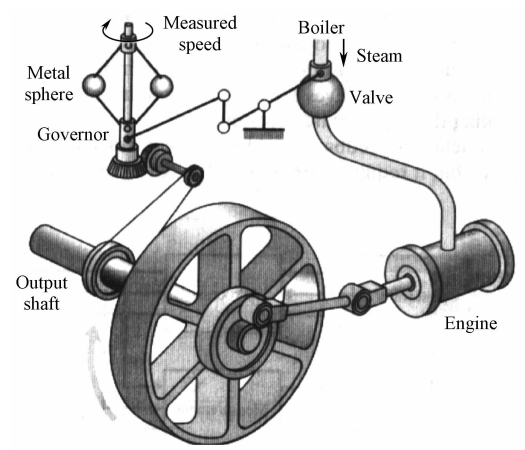


Figure 4. 1. 1 Watt's fly ball governor.

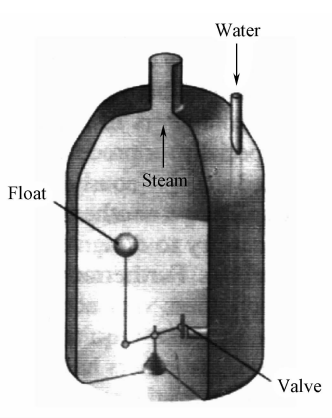


Figure 4. 1. 2 Water-level float regulator.

Prior to World War II, control theory and practice developed in a different manner in the United States and western Europe than in Russia and eastern Europe. A main impetus for the use of feedback in the United States was the development of the telephone system and electronic feedback amplifiers by Bode, Nyquist, and Black at Bell Telephone Laboratories. The frequency domain was used primarily to describe the operation of the feedback amplifiers in terms of bandwidth and other frequency variables. In contrast, the eminent mathematicians and applied mechanicals in the former Soviet Union inspired and dominated the field of control theory. Therefore, the Russian theory tended to utilize a time-domain formulation using differential equations.

A large impetus to the theory and practice of automatic control occurred during World War II when it became necessary to design and construct automatic airplane pilots, gun-positioning systems, radar antenna control systems, and other military systems based on the feedback control approach^[4]. The complexity and expected performance of these military systems necessitated an extension of the available control techniques and fostered interest in control systems and the development of new insights and methods. Prior to 1940, for most cases, the design of control systems was an art involving a trail-and-error approach. During the 1940s, mathematical and analytical methods increased in number and utility, and control engineering became an engineering discipline in its own right.

Frequency-domain techniques continued to dominate the field of control following World War II with the increased use of the Laplace transform and the complex frequency

plane. During the 1950s, the emphasis in control engineering theory was on the development and use of the s -plane methods and, particularly, the root locus approach. Furthermore, during the 1980s, the utilization of digital computers for control components became routine. The technology of these new control elements to perform accurate and rapid calculations was formerly unavailable to control engineers. There are now over 400,000 digital process control computers installed in the United States. These computers are employed especially for process control systems in which many variables are measured and controlled simultaneously by the computer.

With the advent of Sputnik and the space age, another new impetus was imparted to control engineering. It became necessary to design complex, highly accurate control systems for missiles and space probes. Furthermore, the necessity to minimize the weight of satellites and to control them very accurately has spawned the important field of optimal control. Due to these requirements, the time-domain methods developed by Liapunov, Minorsky, and others have met with great interest in the last two decades. Recent theories of optimal control developed by L. S. Pontryagin in the former Soviet Union and R. Bellman in the United States, as well as recent studies of robust systems, have contributed to the interest in time-domain methods. It now is clear that control engineering must consider both the time-domain and the frequency-domain approaches simultaneously in the analysis and design of control systems.

A selected history of control system development is summarized in Table 4.1.1.

Table 4.1.1 Selected Historical Developments of Control Systems.

1769	James Watt's steam engine and governor developed. The Watt steam engine is often used to mark the beginning of the Industrial Revolution in Great Britain. During the Industrial Revolution, great strides were made in the development of mechanization, a technology preceding automation.
1800	Eli Whitney's concept of interchangeable parts manufacturing demonstrated in the production of muskets. Whitney's development is often considered to be the beginning of mass production.
1868	J. C. Maxwell formulates a mathematical model for a governor control of a steam engine.
1913	Henry Ford's mechanized assembly machine introduced for automobile production.
1927	H. W. Bode analyzes feedback amplifier.
1932	H. Nyquist develops a method for analyzing the stability of systems.
1952	Numerical control (NC) developed at Massachusetts Institute of Technology for control of machine-tool axes.
1954	George Devol develops "programmed article transfer," considered to be the first industrial robot design.
1960	First Unimate robot introduced, based on Devol's designs. Unimate installed in 1961 for tending die-casting machines.
1970	State-variable models and optimal control developed.
1980	Robust control system design widely studied.
1990	Export-oriented manufacturing companies emphasize automation.
1994	Feedback control widely used in automobiles. Reliable, robust systems demanded in manufacturing.
1997	First ever autonomous rover vehicle, known as Sojourner, explores the Martian surface.
1998-2003	Advances in micro-and nanotechnology. First intelligent micromachines are developed and functioning nanomachines are created.

4.1.2 Specialized English words

regulator 调节器,调节装置	variables 变量
float regulator mechanics 浮球调节装置,浮阀调节装置	time-domain 时域
water clock 水钟	antenna 天线
water-level 水位	discipline 学科,纪律
steam boiler 蒸汽锅炉	Laplace transform 拉氏变换
pressure-cocker valve 压力锅(安全)阀门	complex frequency plane 频域复平面
fly ball governor 飞球调速器	s-plane s平面
shaft (机器的)轴,井筒	root locus approach 根轨迹法
steam engine 蒸汽机	Sputnik 普尼克(前苏联 1957 年 10 月 4 日发射的第一颗人造卫星)
axis 轴(心)线	probes (航天)探测器,传感器,探针
flyweights 飞行重块,飞行重物	optimal control 最优控制
inlet (液体或气体的)入口	robust system 鲁棒系统
attenuation 衰减,递减	mechanization 机械化
transient oscillations 过渡过程振荡	musket 滑膛枪,毛瑟枪
unstable 不稳定的	interchangeable parts 互换性零(部)件
formulate 构想,规划,确切表达	numerical control (NC) 数值控制,数控
differential equation 微分方程	machine-tool 机床,工具机
parameter 参数	axes 坐标轴(系)
system performance 系统性能,系统表现	robot 机器人
model 模型	die-casting 压铸,模铸
amplifier 放大器	state-variable 状态变量
frequency domain 频(率)域	autonomous 自治的
bandwidth 频宽,带宽	Martian 火星的,火星入
	nanotechnology 纳米技术

4.1.3 Notes

[1]Heron of Alexandria, who lived in the first century A. D. , published a book entitled Pneumatica, which outlined several forms of water-level mechanisms using float regulators. 这是一个典型的复合句,句中有两个定语从句。“Who lived in the first century A. D. ”是“Heron of Alexandria”的定语从句;“which outlined…”是“a book titled Pneumatica”的定语。注意句末的“using floate regulators”为现在分词短语,做“mechanisms”的后置定语。全句可译为“生活在公元一世纪亚历山大城的 Heron 出版了

一本名为“Pneumatica”(《气体力学》)的书,其中记载了好几种采用浮球调节器的水面控制装置。”

[2]J. C. Maxwell formulated a mathematical theory related to control theory using a differential equating model of a governor. 句中“related to control theory”为过去分词短语,做“mathematical theory”的后置定语。而“using a differential equating model of a governor”为现在分词短语,做“control theory”的后置定语。全句可译为“J. C. Maxwell 借助一个蒸汽机调速器的微分方程模型提出了有关控制理论的数学原理。”

[3]Maxwell’s study was concerned with the effect various system parameters had on the system performance. 句中“various system parameters had”为“effect”的定语从句,前面省去了“which”。全句可译为“Maxwell 的研究着眼于各种系统参数对系统性能的影响。”

[4]A large impetus to the theory and practice of automatic control occurred during World War II when it became necessary to design and construct automatic airplane pilots, gun-positioning systems, radar antenna control systems, and other military systems based on the feedback control approach. 本句为并列复合句。注意“when”的词意,为“当时”、“当其时”之意,不是“当……的时候”,表示两个从句并列的时间关系。全句可译为“第二次世界大战时期,自动控制的理论和应用有了重大发展。当时,需要设计和制造基于反馈控制的自动飞机驾驶仪、火炮定位系统、雷达天线控制系统以及其他军事系统。”

4.1.4 Reference Translation

自动控制发展史

将反馈用于控制系统的历史堪称辉煌。最早的反馈控制控制于公元前 300 年至公元前 1 年间出现在希腊开发浮球调节装置的过程中;Ktesibios 的水钟采用了一个浮阀调节器。大约公元前 250 年,Pilon 在他发明的油灯中也装有一个调节器,将燃油面保持在一个恒定的高度上。生活在公元一世纪亚历山大城的 Heron 出版了一本名为“Pneumatica”(《气体力学》)的书,其中记载了好几种采用浮球调节器的水面控制装置。

近代欧洲的第一个反馈系统是荷兰人 Cornelis Drebbel (1572—1633)所发明的一个温度调节装置。

Dennis Papin (1647—1712)于 1681 年发明了首个用于蒸汽锅炉的压力调节器。Papin 的压力调节器是安全调节器的一种,和高压锅的安全阀类似。

第一个用于工业过程的自动反馈控制器公认是 James Watt 于 1769 年发明的用于蒸汽机速度控制的飞球调节器。如图 4.1.1 所示,这个全机械的装置测量输出传动轴的转速,并通过有一定转速的飞球的运动来控制阀门,进而控制进入引擎的蒸汽量。当速度增加时,重球上升,偏离轴线,使阀门关闭。飞球必须从引擎获得能量产生旋转,这导致速度的测量精度不高。

俄国宣称第一个历史性的反馈系统是由 I. Polzunov 于 1765 年发明的水位浮球调节器。这个水位调节系统如图 4. 1. 2 所示。浮球检测水位并控制阀门去关闭锅炉的进水管。

1868 年以前,自动控制系统的发展以直觉发明为主要特点。提高控制系统的精度的努力导致系统瞬态振荡减振过程延长,甚至系统不稳。这样,开发自动控制理论的问题就迫在眉捷。J. C. Maxwell 借助一个蒸汽机调速器的微分方程模型提出了有关控制理论的数学原理。Maxwell 的研究着眼于各种系统参数对系统性能的影响。与此同时,I. A. Vyshnegraadskii 提出了一种调节器的数学理论。

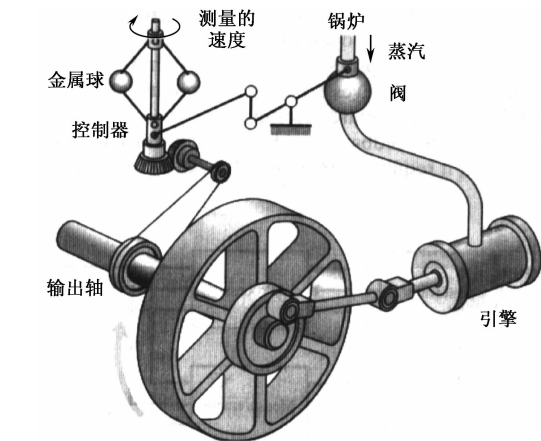


图 4. 1. 1 Watt 的飞球调节器

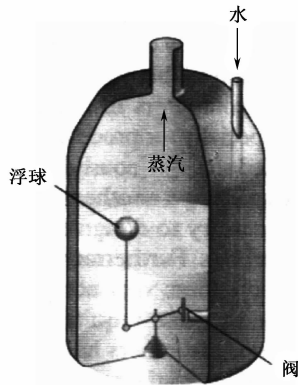


图 4. 1. 2 水位浮球调节器

二战之前,美国和西欧的控制理论和应用的发展方式和俄国及东欧不同。在美国,推动反馈技术应用的主要动力来自贝尔电话实验室的 Bode、Myquist 和 Black 等人对电话系统以及电子反馈放大器的研究,方法上主要采用频域法,用带宽和其他频率变量来描述反馈放大器的运行状态。与此相对,在前苏联,由杰出的数学家和应用机械专家们推动并主宰了控制理论领域。所以俄罗斯的理论倾向于微分方程的时域表达。

第二次世界大战时期,自动控制的理论和应用有了重大发展。当时,需要设计和制造基于反馈控制的自动飞机驾驶仪、火炮定位系统、雷达天线控制系统以及其他军事系统。这些军事系统的复杂性,及对性能的预期都要求开拓更有效的控制技术,促使人们更加关注控制系统,催生了很多新的见解和方法。1940 年之前,控制系统的设计多半还是一门靠反复试验摸索的艺术。到 20 世纪 40 年代,数学和分析方法在数量和实用效果上都大有增长。就这样,控制工程凭着自身的发展成为了一门工程学科。

随着拉氏变换和频域复平面方法应用的发展,频域法在二战时期继续左右着自动控制领域。到了 20 世纪 50 年代,控制工程理论主要的发展和应用是 s 平面法,特别是根轨迹法。到 20 世纪 80 年代,将数字计算机用作控制部件已成平常之事。这种快速而精确的控制手段,是以往的控制工程师们所不具备的。目前,在美国已安装了超过 40 万台的

数字过程控制计算机。这些计算机主要用在那些变量数目很多的过程控制系统中,需要计算机对这些变量进行同步测量和控制。

“普尼克”人造卫星的发射和空间时代的到来,为自动控制技术提供了新的发展动力。为火箭和航天探测器设计复杂而高精度的控制系统已成为必要。更有甚者,为了尽可能减轻卫星的重量,并对卫星进行高精度的控制,催生了最优控制这一重要领域的出现。在过去 20 年中,由 Liapunov、Minorsky 和其他一些人提出的时域方法,满足了人们对这些要求的极大关注。近年由前苏联的 L. S. Pontryagin 和美国的 R. Bellman 所发展的最优控制理论以及对鲁棒系统的研究,都对时域方法做出了贡献。

表 4.1.1 控制系统发展大事选编

1969	Jams Watt 的蒸汽机和调速器问世。瓦特的蒸汽机常常被视为英国工业革命开始的标志。在工业革命过程中,机械化——自动化的先行技术——取得了重大的进展
1800	Eli Whitney 关于互换性零件制造的概念在毛瑟枪的生产中得到验证。这一发展通常被认为是规模化生产的开端
1868	J. C. Maxwell 提出蒸汽机的调速控制器的数学模型
1913	Henry Ford 的机械化生产组装线引入到汽车装配
1927	H. W. Bode 对反馈放大器进行了分析
1932	H. Nyquist 发展出一种系统稳定性的分析方法
1952	麻省理工学院开发出数控(NC)方法,用于机床轴系控制系统
1954	George Devel 提出的“程序化物件转移”方法被认为是第一个工业机器人的设计构思
1960	基于 Devel 设计思想的第一台 Unimate 机器人诞生,并于 1961 年安装用于管理压铸机
1970	提出了状态变量模型和最优控制
1980	鲁棒控制系统设计得到广泛研究
1990	自动化受到出口制造商的重视
1994	反馈控制广泛用到汽车上。制造业要求高可靠的鲁棒系统
1997	名为 Sojourner 的首台自动漫游车车勘查火星表面
1998—2003	微米技术和纳米技术取得进展。首批智能微米机器诞生。功能性纳米机器制成

4. 1. 5 Reading Materials

Modern Control Theory (I)

The desire to control the forces of nature has been with man since early civilizations. Although many examples of control systems exited in early times, it was not until the mid-eighteenth century that several steam operated control devices appeared.

The period beginning about twenty-five years before World War Two saw rapid advances in electronics and especially in circuit theory, aided by the now classical work of Nyquist in the area of stability theory. The advent of the analog computer coupled with advances in electronics saw the beginning of the establishment of control systems as a science. The availability of computers also opened the era of data logging(数据登录), computer control, and the state space of modern method of analysis.

Several factors provided the stimulus(激励, 促进) for the development of modern control theory:

- a. The necessity of dealing with more realistic models of systems.
- b. The shift in emphasis towards optimal control and optimal system design.
- c. The continuing developments in digital computer technology.
- d. The shortcomings of previous approaches.
- e. A recognition of the applicability of well-known methods in other fields of knowledge.

The continuing advances in computer technology have had three principal effects on the controls field. One of these relates to the gigantic supercomputers. The size and class of problems that can now be modeled, analyzed, and controlled are considerably larger than they were when the first of this book was written.

The second impact of computer technology has to do with the proliferation(增长, 繁殖) and wide availability in homes and in the work place. Classical control theory was dominated by graphical methods because at the time that was the only way to solve certain problems. Now every control designer has easy access to powerful computer packages(程序包, 软件包) for system analysis and design. The old graphic methods have not yet disappeared, but have been automated. They survive because of the insight and intuition that they can provide. However, some different techniques are often better suited to a computer. Although a computer can be used to carry out the classical transform-inverse transform(变换-反变换) methods, it is usually more efficient for a computer to integrate differential equations directly.

4.2 The New Generation of Advanced Process Control

4.2.1 Text

Modern control, fuzzy logic, knowledge engineering, and neural networks are the driving forces behind a new architecture of advanced process control.

During the last decade, control scientists and experts have made great efforts to explore the future direction of control theory and its applications. Seeking new technologies for advancing process control is one of the most important challenges the control industry community (control manufacturers and end-users) faces today and tomorrow. Recent advances in computer technology, modern control techniques and machine intelligence have opened a path to the new generation of advanced process control.

The history of process control has followed a conceptual evolution^[1]. Often these concepts have been turned into reality. Yet to a large extent, conventional control is

still the most popular one in industrial applications.

However, due to the increasing complexity of industrial systems, some process control-related technologies, such as modern control, fuzzy sets, expert systems and artificial neural networks are now being applied to the real industrial world and are beginning to show their specific functions in performing controls for complex industrial systems.

Complex systems may be characterized by their main features; they are highly nonlinear, time-variant, large scale and seriously interconnected (i. e. some of the variables may not be linearly independent)^[2]. They may also have incomplete sensory information, system uncertainties (system parameter structure and measurements), and involve difficult human factors^[3].

As a result, the desired control systems for complex industrial processes must have robust and/or adaptive behaviors, in order to adapt to changes of control criteria, system environment and human preference. Obviously, the existing process control strategies can't be effectively applied to complex systems controls^[4].

Development of advanced control for complex systems has been one of the key issues for future industrial control. This article surveys advanced control technologies, and then proposes an integrated architecture for development of a new generation of advanced process control^[5].

Modern Control Techniques

Modern control can be grouped into three major categories:

—State space models, and system identification and estimation are used to estimate or predict system parameters and states, which are related to system internal information being unavailable for sensory measurement^[6].

—Optimal control yields an optimal or suboptimal solution, based on given criteria and system constraints.

—Robust (with fixed controller parameters) and adaptive controls (with self-adjustable parameters and structures) yield satisfactory control behaviors, under system criteria and environment changes^[7].

Precise mathematical descriptions and processing of real control problems are the key measure of modern control. Hence, the precisely quantitative solutions are provided in real-time control. However, it is difficult to deal with system qualitative knowledge, reasoning and human factors using modern control techniques. Further, the applications of modern control require a precise mathematical model and are mainly applicable to linear, time-invariant systems. However, those conditions usually are also not satisfied for complex systems.

Expert Control

Expert control can be applied to complex systems to deal with qualitative knowledge, reasoning and human factors, which are hardly possible to be modeled by mathematical tools. As a result, expert controls are particularly applicable to adaptive control, supervisory control, fault diagnosis, scheduling and planning. In these areas, human knowledge still plays an important role in decision making.

It should, however, be emphasized that the knowledge representation, reasoning and learning in developing expert control for dynamic systems have not been completely solved yet. Providing an expert system environment with the ability to deal with both “qualitative knowledge” and modern control algorithms will be the challenge of expert control development^[8].

Fuzzy Control

Fuzzy control has been considered one of the most powerful strategies for complex systems control, particularly for nonlinear, uncertainty systems. In contrast to the regular knowledge base, the fuzzy relational model and the fuzzy control algorithm can be established based on fuzzy rules, production data and fuzzy identification, and then the system semiquantitative information can be provided^[9].

Note that the accuracy of the fuzzy model and control are based on the size of the fuzzy set determined by the designer. In addition, self-learning can also be implemented by updating the fuzzy relational matrix illustrating system I/O relations^[10].

Similar to modern control, fuzzy modeling and controls can also be applied to system modeling, state estimation and prediction, optimal control, and robust and adaptive controls. But mathematical models are not required, and the approaches can be applied to complex systems with fuzzy knowledge.

Artificial Neural Network Control

An artificial neural network control (ANN) can be considered as a large-scale nonlinear dynamic system consisting of a large number of simple nonlinear processing elements, so-called “neurons”, which are interconnected with each other with adjustable strength^[11]. The major features of ANN are of parallel processing, associative and distributed memory, and learning (learning from example, self-organizing, etc.), which are similar to those evident in biological neural networks^[12].

In contrast to an expert system, a well-trained ANN model can provide both qualitative and quantitative (analogical, digital, or logical) knowledge. ANN has powerful functions in learning and self-organizing, and those properties make ANN much more elegant than expert systems.

It is well known that a great amount of learning, association, adaptation, sensory

information and parallel processing are the major features of human control. These features are distinct from present man-made controllers. However, human control cannot provide precisely quantitative results like man-made controllers.

An ANN controller could be designed as a neural network architecture compromising the features of human control and man-made controllers. However, it should be noted that it is still early in the development of a realistic ANN controller^[13].

The following key problems have to be solved in future R&D:

—Even though some ANN structures have been proposed for dynamic modeling and control, the further development of a generic ANN architecture and its properties for dynamic controls is needed.

—Besides the popular back propagation learning, it is necessary to develop faster and simpler ANN adaptive learning algorithms for dynamic controls.

—Since the special ability of ANN is in combining knowledge, reasoning and self-learning, the coupling of ANN with expert, fuzzy, and modern control will be the way to realize the greatest potential in development of ANN-oriented process control^[14].

Future Direction of DCSs

Modern computer technology has provided an excellent environment for implementing plant-wide computer integrated production. In addition, many software tools are now available in the market for engineers to perform model-based control, expert control, fuzzy logic control and artificial neural network control.

However, if we investigate the real applications in detail, we will find the following facts:

—The existing distributed control systems have made significant improvement in computer technology, but not enough in new control concepts, architectures and strategies. This fact tells us that the movement of advanced process control from academia to applications is still slow.

—In order to develop a new generation of distributed control systems, the research and development based on an integration of the technologies as described in this article is vitally important. Future competition between control manufacturers will not only depend on what advanced computer technology is used, but perhaps depend more heavily on what kind of advanced application software is available for solving complex systems control problems.

—Besides the development for individual tools, such as neural networks, fuzzy control and expert control, etc. the further development of an integrated software environment will be one of the most important tasks for the control industry.

Advanced process control is a multiple discipline. It is necessary to have a strong

team to develop and apply the new technologies. The team needs to have computer experts as well as control experts. To develop a new generation of distributed control systems, those with control expertise will have to have knowledge of modern control, neural networks, artificial intelligence, as well as process fundamentals^[15].

4.2.2 Specialized English words

advanced process control	高级过程控制	adjustable strength	强度可调, 程度可调
reasoning	推理	system uncertainties	系统不确定性
machine intelligence	机器智能	parallel processing	并行处理
linear time invariant systems	线性时不变系统	parameter	参数
fuzzy sets	模糊集	associative and distributed memory	协同与分布式记忆
human knowledge	人类知识, 人工知识	robust	鲁棒性
expert systems	专家系统	R&D	研究与发展, 研发 (Research and Development 之缩写)
knowledge representation, reasoning and learning	知识表示、推理与学习	adaptive	(自)适应的
artificial neural networks	人工智能神经网络	back propagation learning	反向传导学
nonlinear, uncertainty systems	非线性不确定系统	control criteria	控制标准
complex (industrial) systems	复杂(工业)系统	ANN-oriented	面向 ANN 的
fuzzy identification	模糊识别, 模糊辨识	human preference	人工干预, 人为干预
qualitative knowledge	定性知识	compromising	折中, 平衡, 综合
nonlinear	非线性	integrated architecture	综合体系
semi quantitative	半定量的	DCSs	数字控制系统 (Digital Control Systems 之缩写)
time-variant	时变(的)	state space models	状态空间模型
fuzzy relational matrix	模糊关系矩阵	model-based control	基于模型的控制
seriously interconnected	高度相关的, 高度耦合的	system identification and estimation	系统辨识与估计
fuzzy modeling	模糊建模	sensory measurement	传感测量
variables	变量	optimal control	最优控制, 优化控制
neurons	神经细胞, 神经元	academia	学术界
incomplete sensory information	信息感知(的)不完备(性)	suboptimal	次优的
		multiple discipline	综合学科, 交叉学科
		system constraints	系统约束条件
		precise mathematical descriptions	精确

4.2.3 Notes

[1] The history of process control has followed a conceptual evolution. 此句用现在完成时表示“一贯如此”的意思,全句可翻译为“观念的变革造就了过程控制的全部历史。”

[2] Complex systems may be characterized by their main features; they are highly nonlinear, time-variant, large scale and seriously interconnected (i. e. some of the variables may not be linearly independent). 句中“be characterized”和“features”在语意上是重复的,翻译时当然不必死译。全句可翻译为“复杂系统具有以下重要特征:高度非线性、时变性、规模大且高度相关(即有些变量可能不是线性独立的)。”

[3] They may also have incomplete sensory information, system uncertainties (system parameter structure and measurements), and involve difficult human factors. 此句有两个并列的谓语动词:“have”和“involve”。全句可译为“它们还可能具有信息感知的不完备性、系统的不确定性(系统参数结构与测量),以及难以处理的人工因素。”

[4] Obviously, the existing process control strategies can't be effectively applied to complex systems controls. 句中的“can't be effectively applied to…”意为“无法有效地用于……”。全句可翻译为“很明显,现有的过程控制方法很难有效地用于复杂系统的控制。”

[5] This article surveys advanced control technologies, and then proposes an integrated architecture for development of a new generation of advanced process control. 理解此句不难,但应注意两个动词“surveys”和“proposes”的译法。全句可翻译为“本文首先全面介绍各种高级控制技术,然后提出开发新一代高级过程控制的综合架构体系。”

[6] State space models, and system identification and estimation are used to estimate or predict system parameters and states, which are related to system internal information being unavailable for sensory measurement. 句中的“which”引导非限定性定语从句,用来修饰“system parameters”和“states”。从句中的“being unavailable for sensory measurements”为现在分词短语,做“internal information”的定语。全句可翻译为“状态空间模型和系统辨识与评估用来对系统的参数和状态进行评估和预测,这些参数和状态与系统内部信息有关,但无法进行传感测量。”

[7] Robust (with fixed controller parameters) and adaptive controls (with self-adjustable parameters and structures) yield satisfactory control behaviors, under system criteria and environment changes. “under”以下为条件状语从句。此句可译为“当系统准则和环境发生变化时,(控制器参数固定不变情况下的)鲁棒控制和(参数与结构可自适应情况下的)自适应控制都可得到令人满意的控制效果。”

[8] Providing an expert system environment with the ability to deal with both “qualitative knowledge” and modern control algorithms will be the challenge of expert

control development. 句中的短语“providing A with B”意为将“B”提供给“A”,在这里以动名词短语形式做主语。“environment”一词指专家系统体系。全句可翻译为“让专家系统具备处理定性知识和现代控制算法两种能力,是开发专家系统将面临的挑战。”

[9] In contrast to the regular knowledge base, the fuzzy relational model and the fuzzy control algorithm can be established based on fuzzy rules, production data and fuzzy identification, and then the system semi quantitative information can be provided. 句中的“base”与“rules”之意相通,为“规则”之意。“then”引起的是表示结果的从句。全句可翻译为“与常规知识库不同,模糊关系模型和模糊控制算法可以在模糊规则、生产数据及模糊辨识的基础上建立起来,这样就可以提供系统的半定量信息。”

[10] In addition, self-learning can also be implemented by updating the fuzzy relational matrix illustrating system I/O relations. 句中的“illustrating…”为现在分词短语,做后置定语,修饰“matrix”。全句可翻译为“此外,可以通过更新描述系统 I/O 关系的模糊关系矩阵来实现自学习。”

[11] An artificial neural network control (ANN) can be considered as a large-scale nonlinear dynamic system consisting of a large number of simple nonlinear processing elements, so-called “neurons”, which are interconnected with each other with adjustable strength. 句中的“as a system”为主语补足语,“system”后接的“consisting of…”为其后置定语。“which…”则是“neurons”的定语从句。全句可翻译为“人工神经网络(ANN)可以视为由大量称为‘神经元’的结构简单的非线性处理单元所组成的一个大规模非线性动态系统,这些神经元互相交连,且交连程度可调。”

[12] The major features of ANN are of parallel processing, associative and distributed memory, and learning (learning from example, self-organizing, etc.), which are similar to those evident in biological neural networks. 句中的“which”引导定语从句,“which”指代“parallel processing, associative and distributed memory, and learning”,从句中的“those”又带一个定语成分“evident in…”。全句可翻译为“ANN 的主要特点包括并行处理、协同与分布记忆及学习(通过事例学习,自组织)等,这些都和生物神经网络的明显特征相似。”

[13] However, it should be noted that it is still early in the development of a realistic ANN controller. 句中的“early”为形容词,做表语,意为“尚早”。“realistic”意为“实际的、实用的”。全句译为“然而,值得注意的是,对于开发出一个实用 ANN 控制器而言,目前为时尚早。”

[14] Since the special ability of ANN is in combining knowledge, reasoning and self-learning, the coupling of ANN with expert, fuzzy, and modern control will be the way to realize the greatest potential in development of ANN-oriented process control. 这是一个典型的主从复合句,结构不算复杂。句中,由“combining”和“and”组成词组“combining…and…”;由“coupling of”和“with”组成词组“coupling of…with…”。据此可以看出,“the coupling of ANN with expert, fuzzy, and modern control”为主句的主语。

全句可翻译为“由于 ANN 具有把知识、推理和自学习结合起来的独特能力,所以把 ANN 与专家系统、模糊逻辑及现代控制方法整合到一起,将是在开发面向 ANN 的过程控制技术中能最大程度发挥其潜力的一种方法。”

[15] To develop a new generation of distributed control systems, those with control expertise will have to have knowledge of modern control, neural networks, artificial intelligence, as well as process fundamentals. 这是一个简单句。“those”为主语,介词短语“with control expertise”是它的定语。此句可译为“为了开发新一代的分布式控制系统,控制专家们必须具备现代控制、神经网络、人工智能及过程控制基本原理等多方面知识。”

4.2.4 Reference Translation

新一代高级过程控制

现代控制、模糊逻辑、知识工程和神经网络是新一代高级过程控制体系的驱动力量。

在过去十年里,控制科学家和专家们以极大的努力去开拓控制理论的未来方向和应用。为过程控制的发展探寻新的技术是控制业界(控制制造商和用户)今天和明天所面临的最重要的挑战之一。近来在计算机技术、现代控制技术以及机器智能等方面的进展已经打开了一条通往新一代高级过程控制的道路。

观念的变革造就了过程控制的全部历史。这些观念通常可以变为现实。但是在很大程度上,常规控制在工业应用中仍然是最普遍的。

然而,由于工业系统的复杂性不断增加,一些与过程控制有关的技术,例如现代控制、模糊集合、专家系统及人工神经网络正被用于实际工业领域,并开始显示出它们在复杂工业系统中独特的性能。

复杂系统具有以下重要特征:高度非线性、时变性、规模大且高度相关(即有些变量可能不是线性独立的)。它们还可能具有信息感知的不完备性、系统的不确定性(系统参数结构与测量),以及难以处理的人工因素。

于是,复杂工业过程理想的控制系统应该具备鲁棒性以及自适应性的行为模式。很明显,现有的过程控制方法很难有效地用于复杂系统的控制。本文首先全面介绍各种高级控制技术,然后提出开发新一代高级过程控制的综合架构体系。

现代控制技术可以分为三个主要方面:

- 状态空间模型和系统辨识与评估用来对系统的参数和状态进行评估和预测,这些参数和状态与系统内部信息有关,但无法进行传感测量;
- 最优控制可根据给定标准和系统的约束条件获得最优或次优解;
- 当系统准则和环境发生变化时,(控制器参数固定不变情况下的)鲁棒控制和(参数与结构可自调整情况下的)自适应控制都可得到令人满意的控制效果。

实际控制问题的精确数学描述和处理是现代控制技术的关键手段。因此为实时控制提供了精确定量解。但是,用现代控制技术难以处理系统的定性知识、推理以及人为因

素。此外,现代控制技术的要求有一个精确的数学模型,且主要是用于线性时不变系统。然而复杂系统往往很难满足这些条件。

专家控制

专家控制可以用于复杂系统,处理那些很难用数学工具建模的定性知识、推理及人工因素。所以,专家控制特别适合于自适应控制、监控系统、故障诊断、规划调度等领域。在这些领域里,人工知识在决策时仍然会起到重要作用。

让专家系统具备处理定性知识和现代控制算法两种能力,是开发专家系统将面临的挑战。

模糊控制

与常规知识库不同,模糊关系模型和模糊控制算法可以在模糊规则、生产数据及模糊辨识的基础上建立起来,这样就可以提供系统的半定量信息。

此外,可以通过更新描述系统 I/O 关系的模糊关系矩阵来实现自学习。

和现代控制技术相似,模糊建模和模糊控制也可以用于系统建模、状态评估与预测、最优控制及鲁棒与自适应控制,而无需建立数学模型。这些方法还可以通过模糊知识方式用于复杂系统。

人工神经网络控制

人工神经网络(ANN)可以视为由大量称为“神经元”的结构简单的非线性处理单元所组成的一个大规模非线性动态系统,这些神经元互相交连,且交连程度可调。ANN 的主要特点包括并行处理、协同与分布记忆及学习(通过事例学习,自组织)等,这些都和生物神经网络的明显特征相似。

和专家系统不同,一个经过良好训练的 ANN 模型可以提供定性和定量两方面的(模拟、数字或逻辑)知识。ANN 在学习和自组织方面功能强大,这使得 ANN 远比专家系统优良。

很清楚,人工控制的主要特点是大量的学习、协调、适应能力、信息传感和并行处理。这些特点是和目前人类制造的控制器的特点所不同的。但是,人工控制不能像人造的控制器那样提供精确的定量结果。

然而,值得注意的是,对于开发出一个实用 ANN 控制器而言,目前为时尚早。

在未来的研发工作中,必须解决以下关键问题:

——尽管已为动态建模和控制提出了若干 ANN 结构,还需要进一步开发研究动态控制的通用 ANN 体系和性质。

——除了流行的相反传导学习之外,还需要为动态控制开发更快更简单的 ANN 自适应学习算法。

——由于 ANN 具有把知识、推理和自学习结合起来的独特能力,所以把 ANN 与专家系统、模糊逻辑及现代控制方法整合到一起,将是在开发面向 ANN 的过程控制技术中能最大程度发挥其潜力的一种方法。

DCS 的未来方向

现代计算机技术已为实现全厂计算机集成生产创造了极好的环境。此外,市场上已有很多软件工具可供工程师们执行建模控制、专家控制、模糊逻辑控制以及人工神经网络控制。

然而,如果我们深入调查实际应用,就能发现以下事实:

——现在分布式控制系统在计算机技术方面已取得重大进步,但在新的控制理论、控制结构和控制方法等方面的发展还不够。这个事实告诉我们高级过程控制从学术走向应用的进展仍然缓慢。

——为了开发新一代分布式控制系统,以本文所介绍的各种技术集成为基础的研发工作是至关重要的。未来控制制造商之间的竞争不仅取决于用了什么样的先进计算机技术,也许更多地还要取决于为解决复杂系统的控制问题能有什么样的高级应用软件可资应用。

——除了开发具体的控制工具,如神经网络、模糊控制以及专家控制等,进一步发展一个集成的软件环境将是控制业最重要的任务之一。

高级过程控制是一个多学科领域,必须拥有一支强有力的团队来开发和应用新技术。这个团队需要有计算机专家,还要有控制专家。为了开发新一代的分布式控制系统,控制专家们必须具备现代控制、神经网络、人工智能及过程控制基本原理等多方面知识。

4.2.5 Reading Materials

Modern Control Theory (II)

The third major impact of computers is that they are now so commonly used as just another component in the control system. Their cost, size and reliability make it possible to use them routinely in many systems. This means that the discrete-time and digital system control now deserves much more attention than it did in the past.

Finally, the refinement(进步) of the chip and related computer development has created an explosion(激增) in computational capability and computer-control devices. This has led to many innovative methods in manufacturing methods, such as computer-aided design and manufacturing(计算机辅助设计与制造), and the possibility of unprecedented increases in industrial productivity via the use of computer-controlled machinery, manipulators and robotics(机器人应用).

According to the encyclopedia American(美利坚百科全书), a system is “an aggregation or assemblage(集合体, 组合体) of things so combined by nature or man as to form an integral and complex whole”. Mathematical systems theory is the study, of the interruptions and behavior of such an assemblage of “things” when subjected to certain conditions or inputs. The abstract nature of systems theory is due to the fact that it is concerned with mathematical properties rather than the physical form of the

constituent parts.

Control theory is more often concerned with physical applications. A control system is considered to be any system which exists for the purpose of regulating or controlling the flow of energy, information, money, or other quantities in some desired fashion. In more general terms, a control system is an interconnection of many components or functional units in such a way to produce a desired result. In this book, control theory is assumed to encompass all questions related to design and analysis of systems.

The design of any engineering system or component requires a specification of the job to be done, or of certain essential properties the system must have. This specification should be fairly precise; it is useful to know when a given design is just good enough for the job at hand since better quality almost invariably results in more complex, difficult, and expensive designs. Automatic control systems are no exception.

4.3 Feedback Fundamentals

4.3.1 Text

Introduction

Fundamental properties of feedback systems will be investigated in this Chapter. We begin in Section 1.2 by discussing the basic feedback loop and typical requirements. This includes the ability to follow reference signals, effects of load disturbances and measurement noise and the effects of process variations. It turns out that these properties can be captured by a set of six transfer functions, called the Gang of Six. These transfer functions are introduced in Section 1.3. For systems where the feedback is restricted to operate on the error signal the properties are characterized by a subset of four transfer functions, called the Gang of Four^[1]. Properties of systems with error feedback and the more general feedback configuration with two degrees of freedom are also discussed in Section 1.3. It is shown that it is important to consider all transfer functions of the Gang of Six when evaluating a control system. Another interesting observation is that for systems with two degrees of freedom the problem of response to load disturbances can be treated separately. This gives a natural separation of the design problem into a design of a feedback and a feedforward system. The feedback handles process uncertainties and disturbances and the feedforward gives the desired response to reference signals. Attenuation of disturbances are discussed in Section 1.4 where it is demonstrated that process disturbances can be attenuated by feedback but that feedback also feeds measurement noise into the system. It turns out that the sensitivity function which belongs to the Gang of Four gives a nice characterization of disturbance

attenuation. The effects of process variations are discussed in Section 1. 5. It is shown that their effects are well described by the sensitivity function and the complementary sensitivity function. The analysis also gives a good explanation for the fact that control systems can be designed based on simplified models. When discussing process variations it is natural to investigate when two processes are similar from the point of view of control. This important nontrivial problem is discussed in Section 1. 6. Section 1. 7 is devoted to a detailed treatment of the sensitivity functions. This leads to a deeper understanding of attenuation of disturbances and effects of process variations. A fundamental result of Bode which gives insight into fundamental limitations of feedback is also derived. This result shows that disturbances of some frequencies can be attenuated only if disturbances of other frequencies are amplified. Tracking of reference signals are investigated in Section 1. 8. Particular emphasis is given to precise tracking of low frequency signals. Because of the richness of control systems the emphasis on different issues varies from field to field. This is illustrated in Section 1. 10 where we discuss the classical problem of design of feedback amplifiers.

The Basic Feedback Loop

A block diagram of a basic feedback loop is shown in Figure 4. 3. 1. The system loop is composed of two components, the process P and the controller. The controller has two blocks the feedback block C and the feedforward block F . There are two disturbances acting on the process, the load disturbance d and the measurement noise n . The load disturbance represents disturbances that drive the process away from its desired behavior. The process variable x is the real physical variable that we want to control. Control is based on the measured signal y , where the measurements are corrupted by measurement noise n . Information about the process variable x is thus distorted by the measurement noise. The process is influenced by the controller via the control variable u . The process is thus a system with three inputs and one output. The inputs are: the control variable u , the load disturbance d and the measurement noise n . The output is the measured signal. The controller is a system with two inputs and one output. The inputs are the measured signal y and the reference signal r and the output is

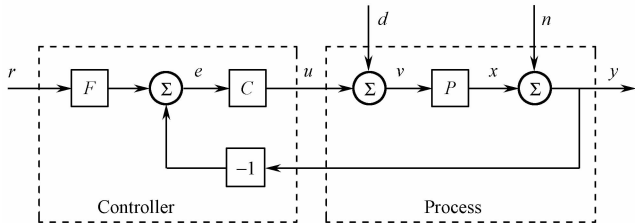


Figure 4. 3. 1 Block diagram of a basic feedback loop.

the control signal u . Note that the control signal u is an input to the process and the

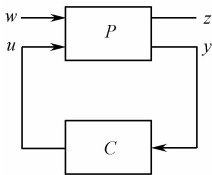


Figure 4.3.2 An abstract representation of the system in Figure 4.3.1. The input u represents the control signal and the input w represents the reference r , the load disturbance d and the measurement noise n . The output y is the measured variables and z are internal variables that are of interest.

output of the controller and that the measured signal is the output of the process and an input to the controller. In Figure 4.3.1 the load disturbance was assumed to act on the process input. This is a simplification, in reality the disturbance can enter the process in many different ways. To avoid making the presentation unnecessarily complicated we will use the simple representation in Figure 4.3.2. This captures the essence

and it can easily be modified if it is known precisely how disturbances enter the system.

More Abstract Representations

The block diagrams themselves are substantial abstractions but higher abstractions are sometimes useful. The system in Figure 4.3.1 can be represented by only two blocks as shown in Figure 4.3.2. There are two types of inputs, the control u , which can be manipulated and the disturbances $w = (r, d, n)$, which represents external influences on the closed loop systems. The outputs are also of two types, the measured signal y and other interesting signals $z = (e, v, x)$. The representation in Figure 4.3.2 allows many control variables and many measured variables, but it shows less of the system structure than Figure 4.3.1. This representation can be used even when there are many input signals and many output signals. Representation with a higher level of abstraction are useful for the development of theory because they make it possible to focus on fundamentals and to solve general problems with a wide range of applications. Care must, however, be exercised to maintain the coupling to the real world control problems we intend to solve.

Disturbances

Attenuation of load disturbances is often a primary goal for control. This is particularly the case when controlling processes that run in steady state^[3]. Load disturbances are typically dominated by low frequencies. Consider for example the cruise control system for a car, where the disturbances are the gravity forces caused by changes of the slope of the road. These disturbances vary slowly because the slope changes slowly when you drive along a road. Step signals or ramp signals are commonly used as prototypes for load disturbances.

Measurement noise corrupts the information about the process variable that the sensors delivers. Measurement noise typically has high frequencies. The average value of the noise is typically zero. If this was not the case the sensor will give very misleading information about the process and it would not be possible to control it well. There may also be dynamics in the sensor. Several sensors are often used. A common situation is that very accurate values may be obtained with sensors with slow dynamics and that rapid but less accurate information can be obtained from other sensors^[4].

Actuation

The process is influenced by actuators which typically are valves, motors, that are driven electrically, pneumatically, or hydraulically. There are often local feedback loops and the control signals can also be the reference variables for these loops. A typical case is a flow loop where a valve is controlled by measuring the flow. If the feedback loop for controlling the flow is fast we can consider the set point of this loop which is the flow as the control variable^[5]. In such cases the use of local feedback loops can thus simplify the system significantly. When the dynamics of the actuators is significant it is convenient to lump them with the dynamics of the process. There are cases where the dynamics of the actuator dominates process dynamics.

Design Issues

Many issues have to be considered in analysis and design of control systems. Basic requirements are

- Stability
- Ability to follow reference signals
- Reduction of effects of load disturbances
- Reduction of effects of measurement noise
- Reduction of effects of model uncertainties

4.3.2 Specialized English Words

transfer function	传递函数	step signals	阶跃信号
error signals	误差信号	ramp signals	斜坡信号
disturbances	扰动,干扰,失调	actuator	调节器,执行机构
degree of freedoms	自由度	valves	阀,电子管
feed ward	前馈	pneumatically	通过气动
attenuation	衰减,减弱,阻压	hydraulically	通过液压(水、油等)
nontrivial	并非不重要的	lump with	将……和……归并到一起
cruise	稳速(行驶)		

4.3.3 Notes

[1] For systems where the feedback is restricted to operate on the error signal the properties are characterized by a subset of four transfer functions, called the Gang of Four. 这个复合句的基本结构并不复杂。主句为“the properties are characterized by a subset of four transfer functions, called the Gang of Four.”介词短语“For systems”用做主句的状语,而“systems”带有一个“where”引导的定语从句。全句可译为“对于那些靠基于误差信号的反馈而运行的系统来说,其基本特征可以用一个由4个传递函数组成的称为‘四(传函)组合’的子集来表示。”

[2] Consider for example the cruise control system for a car, where the disturbances are the gravity forces caused by changes of the slope of the road. 在这个主从复合句中,“where”引起一定语从句,修饰“the cruise control system for a car”,而从句中的表语“gravity forces”有一个做后置定语的去分词短语“caused by...”。全句可译为“让我们考虑控制汽车稳速行驶的例子,这里干扰是由道路坡度变化引起的重力变化。”

[3] This is particularly the case when controlling processes that run in steady state. 这个句子的主句为“This is particularly the case”。“when controlling processes that run in steady state”为状语成分,“when”之后省略了“you are”,“that run in steady state”为“processes”的定语从句。全句可译为“控制的基本目标往往是衰减干扰,在控制稳态运行的过程中尤其如此。”

[4] A common situation is that very accurate values may be obtained with sensors with slow dynamics and that rapid but less accurate information can be obtained from other sensors. 句中两“that”引起的是并列的表语从句。“values”和“information”均指传感输出结果,在同一句中同一对象常用不同词语表达,这是英文行文特点之一。这和中文表达方式不同,中文中同一事物,特别是科技名词更愿前后统一表述,以免产生歧义,这是阅读英文资料时值得注意的。本句可译为“常见的情况是,动态特性慢的传感器的精度非常高,而动态特性快的传感器则精度较差但速度较快。”

[5] If the feedback loop for controlling the flow is fast, we can consider the set point of this loop which is the flow as the control variable. 这是一个主从复合句,“if”引起条件从句,主句为“we can consider...as...”。从句中的宾语“set point of this loop”又带一定语从句“which is the flow”。全句译为“如果流量控制的反馈回路速度够快的话,我们可以把这个回路的设定点,即流量,作为控制变量。”

4.3.4 Reference Translation

反馈基本原理

引言

本章将讨论反馈系统的基本性质。从1.2节开始,将讨论基本反馈回路及典型要求,

包括对参考信号的跟随能力、负载扰动的影响、噪声测量以及过程变化产生的影响。可以得出结论,这些性质可以通过一个由 6 个传递函数组成的函数组来表达,称为“六(传函)组合”。“六组合”将在 1.3 节中介绍。对于那些靠基于误差信号的反馈而运行的系统来说,其基本特征可以用一个由 4 个传递函数组成的称为“四(传函)组合”的子集来表示。1.3 节还将讨论带误差反馈的系统的性质以及具有两个自由度的更为一般的反馈结构。本章指明,在评估一个控制系统时,考虑“六组合”中所有的 6 个传递函数是非常重要的。另一个有趣的结论是,对于两个自由度的系统而言,对负载扰动的响应问题可以单独考虑。这就能自然而然地将设计问题分成反馈系统和前馈系统。反馈处理过程的不确定性和干扰,前馈则对参考信号给出理想的响应。1.4 节讨论干扰的衰减,指出反馈可以衰减过程干扰,但也会将测量噪声引入系统。本节还将证明“四组合”中的灵敏度函数很好地描述了干扰衰减的特征。1.5 节讨论的是过程变化带来的影响,证明灵敏度函数和互补灵敏度函数能很好地描述过程变化所造成的影响。通过分析,清楚地说明了可以在简单模型的基础上进行控制系统的设计。在讨论过程变化时,从控制的观点对两个相似的系统去进行考察是很自然的。这一重要问题将在 1.6 节中讨论。1.7 节详细分析灵敏度函数,更深入地了解干扰衰减及过程变化的影响。本节还对伯德公式的基本结论进行了推导,这一基本结论能深入了解反馈的基本限制。这一结论表明,某些频率的干扰只有在另一些频率的干扰被放大时才会衰减。1.8 节对参考信号的跟随问题进行分析,重点放在低频信号的精确跟随。由于控制系统的多样性,不同领域所强调的问题是不同。这将在讨论反馈放大器设计的典型问题的 1.10 节中予以说明。

基本反馈回路

图 4.3.1 所示为基本反馈回路的框图。系统回路由过程 P 和控制器两部分组成。

控制器由两框组成,反馈框 C 和前馈框 F 。作用到过程的干扰有两个,负载干扰 d 和测量噪声 n 。负载干扰代表驱使过程偏离理想行为的那些干扰。过程变量 x 是我们所要控制的真正的物理变量。控制是以所测量的信号 y 为基础的,而测量会受到测量噪声

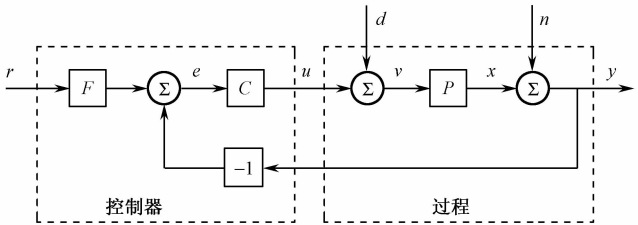


图 4.3.1 基本反馈回路框图

n 扰动。过程变量 x 的信息因而而受到测量噪声的影响。过程又经控制变量 u 受到控制器的影响。所以本系统是一个有着三个输入一个输出的系统。输入为控制变量 u 、负载干扰 d 以及测量噪声 n 。输出为测量信号。控制器是一个有着两个输入一个输出的系统。输入是测量信号 y 和参考信号 r ,而输出信号为控制信号 u 。注意,控制信号 u 是过程的输入和控制器的输出;而测量信号是过程的输出及控制器的输入。图 4.3.1 中,负载干扰假定作用在过程输入处。这是一种简化。在实际中,干扰可以以多种不同方式进入系统。为了避免系统不必要地被复杂化,我们在图 4.3.2 中采用了一种简化的表示。它表达了

实质性的东西。如果精确了解干扰进入系统的途径,本图是很容易修改的。

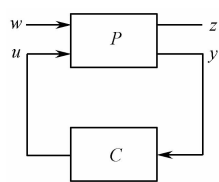


图 4.3.2 图 4.3.1 中所示系统的一种抽象表示。
输入 u 表示控制信号,输入 w 表示参考值 r 、负载干扰 d 及测量噪声 n 。输出 y 是被测变量, z 是我们感兴趣的内部变量

更抽象的表示形式

框图本身是很抽象的,但有时更抽象的东西也很有用。图 4.3.1 中的系统可以仅用两个框来表示,如图 4.3.2 所示。有两种输入:可以控制的控制变量 u 和代表作用在闭环系统上的外来影响的干扰 $w = (r, d, n)$ 。输出也有两种:测量信号 y 及其他有意义的信号 $z = (e, u, x)$ 。图 4.3.2 的表示中可以有多个控制变量和多个测量变量,但却比图 4.3.1 所显示

的系统结构要更简单。这种表示在输入输出信号很多的情况下也可以使用。更高级的抽象表示对于理论研究是有益的,因为这样有可能把焦点汇聚到基本的原则问题上去解决大范围应用的一般性问题。然而要注意的是,理论必须与要解决的实际控制问题结合起来。

干扰

控制的基本目标往往是衰减干扰,在控制稳态运行的过程中尤其如此。负载干扰具有低频的典型特点。让我们考虑控制汽车稳速行驶的例子,这里干扰是由道路坡度变化引起的重力变化。这些干扰变化非常缓慢,因为当你沿着一条路开车时,坡度的变化是平缓的。阶跃信号或斜坡信号常用做负载干扰的典型波形。

测量噪声扰乱了传感器发出的过程变量信息。测量噪声具有高频的典型特性。噪声平均值常常为 0。如果情况不是这样,传感器给出的关于过程的信息会产生严重误导,这就不可能实施正确的控制。传感器本身也会具有不同的动态特征,通常会同时使用若干个传感器。常见的情况是,动态特性慢的传感器的精度非常高,而动态特性快的传感器则精度较差但速度较快。

执行(机构)

过程会受到执行机构的影响。典型的执行机构为阀和电机,由电、气、液压驱动,常常形成局部反馈回路。控制信号也可以用做这些回路的参考变量。一个典型的例子是流量回路,通过测量流量来控制阀门。如果流量控制的反馈回路速度够快的话,我们可以把这个回路的设定,即流量,作为控制变量。在这类例子中,这样应用反馈回路可以大大简化系统。当执行机构的动态特性良好时,可以很方便地将它和过程的动态特性合到一起来考虑。也有执行机构主宰了过程动态特性的例子。

设计问题

进行控制系统的分析和设计时,有很多问题需要考虑,其基本要求包括:

- 稳定性
- 跟随参考信号的能力

- 减少负载干扰的影响
- 减少测量干扰的影响
- 降低模型的不确定性的影响

4.3.5 Reading Materials

Feedback Fundamentals

The possibility of instabilities (不稳定) is the primary drawback of feedback. Avoiding instability is thus a primary goal. It is also desirable that the process variable follows the reference signal faithfully. The system should also be able to reduce the effect of load disturbances. Measurement noise is injected into the system by the feedback. This is unavoidable but it is essential that not too much noise is injected. It must also be considered that the models used to design the control systems are inaccurate. The properties of the process may also change. The control system should be able to cope with (对付) moderate (适度的, 一定的) changes. The focus on different abilities vary with the application. In process control the major emphasis is often on attenuation of load disturbances, while the ability to follow reference signals is the primary concern in motion control systems.

The feedback loop in Figure 4.3.1 is influenced by three external signals, the reference r , the load disturbance d and the measurement noise n .

There are at least three signals x , y and u that are of great interest for control. This means that there are nine relations between the input and the output signals. Since the system is linear these relations can be expressed in terms of the transfer functions. Let X , Y , U , D , N , R be the Laplace transforms of x , y , u , d , n , r , respectively. The following relations are obtained from the block diagram in Figure 4.3.1.

4.4 Frequency Response Methods

4.4.1 Text

Introduction

In Preceding chapters, the response and performance of a system have been described in terms of the complex frequency variable s and the location of the poles and zeros on the s -plane. A very practical and important alternative approach to the analysis and design of a system in the frequency response method.

The frequency response of a system is defined as the steady-state response of the system to a sinusoidal input signal. The sinusoid is unique input signal, and the resulting output signal for a linear system, as well as signals through the system, is

sinusoidal in the steady state; it differs from the input waveform only in amplitude and phase angle.

For example , consider the system $Y(s) = T(s)R(s)$ with $r(t) = A\sin\omega t$. We have

$$R(s) = \frac{A\omega}{s^2 + \omega^2}$$

and

$$T(s) = \frac{m(s)}{q(s)} = \frac{m(s)}{\prod_{i=1}^n (s + p_i)}$$

where p_i are assumed to be distinct poles. Then, in partial fraction form , we have

$$Y(s) = \frac{k_1}{s + p_1} + \dots + \frac{k_n}{s + p_n} + \frac{as + \beta}{s^2 + \omega^2}$$

Taking the inverse Laplace transform yields

$$y(t) = k_1 e^{-p_1 t} + \dots + k_n e^{-p_n t} + L^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\}$$

where α and β are constants which are problem dependent . If the system is stable, then all p_i have positive nonzero real parts and

$$\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} \mathcal{L}^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\}$$

since each exponential term $k_i e^{-p_i t}$ decays to zero as $t \rightarrow \infty$.

In the limit for $y(t)$, we obtain, for $t \rightarrow \infty$ (the steady state) ,

$$\begin{aligned} y(t) &= \mathcal{L}^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\} \\ &= \frac{1}{\omega} |A\omega T(j\omega)| \sin(\omega t + \phi) \\ &= A |T(j\omega)| \sin(\omega t + \phi) \end{aligned} \tag{4.4.1}$$

where $\phi = \angle T(j\omega)$.

Thus, the steady-state output signal depends only on the magnitude and of $T(j\omega)$ at a specific frequency ω . Notice that the steady-state response, as described in Equation (4.4.1), is true only for stable systems, $T(s)$.

One advantage of the frequency response method is the ready available sinusoid test signals for various ranges of frequencies and amplitudes. Thus, the experimental determination of the system's frequency response is easily accomplished; it is the most reliable and uncomplicated method for the experimental analysis of a system. Often, as we shall find in Section 1.4, the unknown transfer function of a system can be deduced from the experimentally determined frequency response of a system. Furthermore, the design of a system in the frequency domain provides the designer with control of the bandwidth of a system, as well as some measure of the response of the system to

undesired noise and disturbance^[1].

A second advantage of the frequency response method is that the transfer function describing the sinusoidal steady-state behavior of a system can be obtained by replacing s with $j\omega$ in the system transfer function $T(s)$ ^[2]. The transfer function representing the sinusoidal steady-state behavior of a system is then a function of the complex variable $j\omega$ and is itself a complex function $T(j\omega)$ that possesses a magnitude and phase angle. The magnitude and phase angle of $T(j\omega)$ are readily represented by graphical plots that provide significant insight into the analysis and design of control systems.

The basic disadvantage of the frequency response method for analysis and design is the indirect link between the frequency and the time domain. Direct correlations between the frequency response and the corresponding transient response characteristics are somewhat tenuous, and in practice the frequency response characteristic is adjusted by using various design criteria that will normally result in a satisfactory transient response.

The Laplace transform pair was given in Section 2.4; it is written as

$$F(s) = \mathcal{L} \{ f(t) \} = \int_0^{\infty} f(t) e^{-st} dt \quad (4.4.2)$$

and

$$f(t) = \mathcal{L}^{-1} \{ F(s) \} = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} F(s) e^{st} ds \quad (4.4.3)$$

where the complex variable $s = \sigma + j\omega$. Similarly, the Fourier transfer pair is written as

$$F(j\omega) = \mathcal{F} \{ f(t) \} = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \quad (4.4.4)$$

and

$$f(t) = \mathcal{F}^{-1} \{ F(j\omega) \} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(j\omega) e^{j\omega t} d\omega \quad (4.4.5)$$

The Fourier transform exists for $f(t)$ when

$$\int_{-\infty}^{+\infty} |f(t)| dt < \infty$$

The Fourier and Laplace transforms are closely related, as we can see by examining Equations (4.4.2) and (4.4.4). When the function $f(t)$ is defined only for $t \geq 0$, as is often the case, the lower limits on the integrals are the same^[3]. Then we note that the two equations differ only in the complex variable. Thus, if the Laplace transform of a function $f_1(t)$ is known to be $F_1(s)$, we can obtain the Fourier transform of this same time function $F_1(j\omega)$ by setting $s = j\omega$ in $F_1(s)$.

Again we might ask, since the Fourier and Laplace transforms are so closely related, why can't we always use the Laplace transform? Why use the Fourier transform at all? The Laplace transform permits us to investigate the s -plane location of the poles and zeros of a transfer $T(s)$, as in Chapter 7. However, the frequency response method

allows us to consider the transfer function $T(j\omega)$ and to concern ourselves with the amplitude and phase characteristics of the system. This ability to investigate and represent the character of a system by amplitude, phase equations, and curves is an advantage for the analysis and design of control systems.

If we consider the frequency response of the closed-loop system, we might have an input $r(t)$ that has a Fourier transform in the frequency domain as follows:

$$R(j\omega) = \int_{-\infty}^{+\infty} r(t) e^{-j\omega t} dt$$

Then the output frequency response of a single-loop control system can be obtained by substituting $s = j\omega$, in the closed-loop system relationship, $Y(s) = T(s)R(s)$, so that we have

$$Y(j\omega) = T(j\omega)R(j\omega) = \frac{G(j\omega)}{1 + G(j\omega)H(j\omega)}R(j\omega) \tag{4.4.6}$$

Utilizing the inverse Fourier transform, the output transient response would be

$$y(t) = \mathcal{F}^{-1} \{ Y(j\omega) \} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} Y(j\omega) e^{j\omega t} d\omega \tag{4.4.7}$$

However, it is usually quite difficult to evaluate this inverse transform integral for all but the simplest systems, and a graphical integration may be used. Alternatively, as we will note in succeeding sections, several measures of the transient response can be related to the frequency characteristics and utilized for design purposes.

Frequency Response Plots

The transfer function of a system, $G(s)$, can be described in the frequency domain by the relation

$$G(j\omega) = G(s) \big|_{s=j\omega} = R(\omega) + jX(\omega) \tag{4.4.8}$$

where $R(\omega) = \text{Re}[G(j\omega)]$ and $X(\omega) = \text{Im}[G(j\omega)]$.

See the MCS website for a review of complex numbers.

Alternatively, the transfer function can be represented by a magnitude $|G(j\omega)|$ and a phase $\phi(j\omega)$ as

$$G(j\omega) = |G(j\omega)| e^{j\phi(j\omega)} = |G(j\omega)| \angle \phi(\omega) \tag{4.4.9}$$

where $\phi(\omega) = \arctan \frac{X(\omega)}{R(\omega)}$, $|G(\omega)|^2 = [R(\omega)]^2 + [X(\omega)]^2$.

The graphical representation of the frequency response of the system $G(j\omega)$ can utilize either Equation (4.4.8) or Equation (4.4.9). The polar plot representation of the frequency response is obtained by using Equation (4.4.8). The coordinates of the polar plot are the real and imaginary parts of $G(j\omega)$, as shown in Figure 4.4.1. An example of a polar

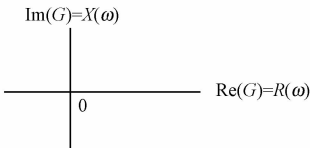


Figure 4.4.1 The polar plane.

plot will illustrate this approach.

Example 4.4.1 Frequency response of an RC filter

A simple RC filter is shown in Figure 4.4.2. The transfer function of this filter is

$$G(s) = \frac{V_2(s)}{V_1(s)} = \frac{1}{RSc + 1} \tag{4.4.10}$$

and the sinusoidal steady-state transfer function is

$$G(j\omega) = \frac{1}{j\omega(RC) + 1} = \frac{1}{j(\omega/\omega_1) + 1} \tag{4.4.11}$$

where $\omega_1 = \frac{1}{RC}$.

Then the polar plot is obtained from the relation

$$\begin{aligned} G(j\omega) &= R(\omega) + jX(\omega) \\ &= \frac{1 - j(\omega/\omega_1)}{(\omega/\omega_1)^2 + 1} \\ &= \frac{1}{1 + (\omega/\omega_1)^2} - \frac{j(\omega/\omega_1)}{1 + (\omega/\omega_1)^2} \end{aligned} \tag{4.4.12}$$

The first step is to determine $R(\omega)$ and $X(\omega)$ at the two frequencies, $\omega = 0$ and $\omega = \infty$. At $\omega = 0$, we have $R(\omega) = 1$ and $X(\omega) = 0$. At $\omega = \infty$, we have $R(\omega) = 0$ and $X(\omega) = 0$. These two points are shown in Figure 4.4.3. The locus of the real and imaginary parts is also shown in Figure 4.4.3 and is easily shown to be a circle with the center at $(1/2, 0)$. When $\omega = \omega_1$, the real and imaginary parts are equal, and the angle $\phi(\omega) = 45^\circ$. The polar plot can also be readily obtained from Equation (4.4.9) as

$$G(j\omega) = |G(\omega)| + \angle \phi(\omega) \tag{4.4.13}$$

where $|G(\omega)| = \frac{1}{[(\omega/\omega_1)^2 + 1]^{1/2}}$ and $\phi(\omega) = -\arctan(\omega/\omega_1)$.

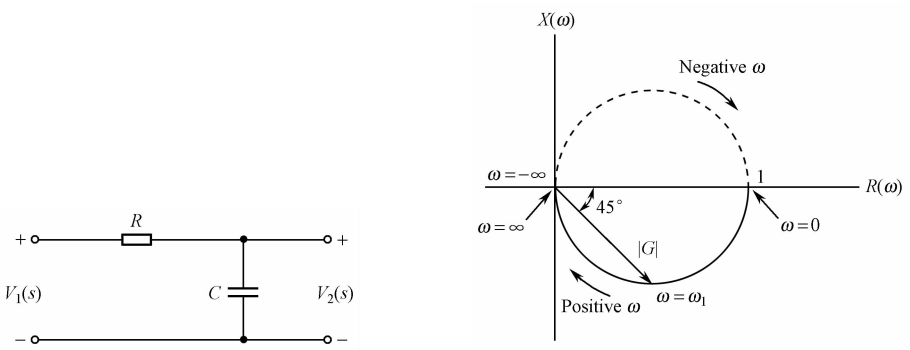


Figure 4.4.2 An RC filter.

Figure 4.4.3 Polar plot for filter.

Hence, when $\omega = \omega_1$, the magnitude is $|G(\omega)| = \frac{1}{\sqrt{2}}$ and the phase $\phi(\omega_1) = -45^\circ$.

Also, when ω approaches $+\infty$, we have $|G(\omega)| \rightarrow 0$ and $\phi(\omega_1) = -90^\circ$. Similarly, when $\omega = 0$, we have $|G(\omega)| = 1$, $\phi(\omega_1) = 0$.

4.4.2 Specialized English Words

frequency response	频率响应	deduce	推导
complex frequency variables	频域复变量,复频变量	frequency domain	频域
poles	极点	correlation	相互关系,相关
zeros	零点	tenuous	微弱的,脆弱的
s-plane	s-平面	transient response	瞬态响应,暂态响应
steady-state response	稳态响应	criteria	标准,判据
sinusoidal	正弦(的)	Fourier transform	傅里叶变换,傅氏变换
amplitude	幅值,幅度	polar plots	极图
phase angle	相位角	coordinates	坐标(系)
partial fraction form	部分分式形式	filter	滤波器
exponential term	指数项	locus	(根)轨迹
magnitude	幅度大小		

4.4.3 Notes

[1] Furthermore, the design of a system in the frequency domain provides the designer with control of the bandwidth of a system, as well as some measure of the response of the system to undesired noise and disturbance. 结构上看,这是一个简单句。“provide ...A with B...”这个短语意为“为 A 提供 B”。本句中“with”有两个介词宾语“control of”和“some measure of”。全句可译为“同时,在频域内进行系统设计,设计者能对系统带宽进行控制,还能对系统对噪声和干扰作出的响应进行衡量。”

[2] A second advantage of the frequency response method is that the transfer function describing the sinusoidal steady-state behavior of a system can be obtained by replacing s with $j\omega$ in the system transfer function $T(s)$. 这是一个系表结构的主从复合句。“that”以下是一个表语从句。从句主语为“transfer function”。“describing... of a system”为现在分词短语,做“transfer function”的后置定语,而“replacing... $T(s)$ ”为动名词加宾语结构,做“by”的介词宾语,一起构成介词短语结构,用做从句的状语。全句可译为“频率响应法的另一个优点是,把系统传递函数 $T(s)$ 中的 s 换成 $j\omega$,就可以得到描述系统正弦稳态响应的传递函数。”

[3] When the function $f(t)$ is defined only for $t \geq 0$, as is often the case, the lower limits on the integrals are the same. 这是一个复合句。“the lower limits on the integrals

are the same”为主句。“the function $f(t)$ is defined only for $t \geq 0$ ”为条件状语从句;注意,“as is often the case”为从句的条件状语的状语从句,“as”后面省去了“it”。全句可译为“当函数 $f(t)$ 仅定义在 $t \geq 0$ 时(这是一般情况),两者积分下限相同。”

4.4.4 Reference Translation

频率响应法

引言

在前面的章节中,采用了复频变量 s 及 s 平面上的极点和零点对系统的响应和性能进行了描述。另一种非常实用和重要的系统分析和设计方法就是频率响应法。

系统的频率响应定义为系统对正弦输入信号的稳态响应。正弦信号是单一的输入信号,而线性系统的作为结果的输出信号,以及系统内部的信号,是稳态正弦信号。和输入波形相比,频率相同,只是幅值和相位角不一样。

例如,考虑系统 $Y(s) = T(s)R(s)$, 且 $r(t) = A\sin\omega t$ 。我们有

$$R(s) = \frac{A\omega}{s^2 + \omega^2}$$

及

$$T(s) = \frac{m(s)}{q(s)} = \frac{m(s)}{\prod_{i=1}^n (s + p_i)}$$

其中, p_i 为离散极点。那么,用部分分式的形式来表示,我们有

$$Y(s) = \frac{k_1}{s + p_1} + \dots + \frac{k_n}{s + p_n} + \frac{as + \beta}{s^2 + \omega^2}$$

取拉氏反变换,得

$$y(t) = k_1 e^{-p_1 t} + \dots + k_n e^{-p_n t} + L^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\}$$

其中 α 和 β 是与问题相关的常数。如果系统是稳定的,则所有 p_i 具有非零正实部,而且由于每个指数项, $k_i e^{-p_i t}$ 当 $t \rightarrow \infty$ 时趋于 0,所以

$$\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} L^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\}$$

对于 $y(t)$ 的极限,当 $t \rightarrow \infty$ (稳态)时,我们有

$$\begin{aligned} y(t) &= L^{-1} \left\{ \frac{as + \beta}{s^2 + \omega^2} \right\} \\ &= \frac{1}{\omega} |A\omega T(j\omega)| \sin(\omega t + \phi) \\ &= A |T(j\omega)| \sin(\omega t + \phi) \end{aligned} \tag{4.4.1}$$

其中 $\phi = \angle T(j\omega)$ 。

所以,稳态输出信号仅取决于 $T(j\omega)$ 在一定频率 ω 下的幅值和相位。注意式(4.4.1)

所描述的稳态响应仅对稳态系统 $T(s)$ 才成立。

频率响应法的一个优点是各种频率和幅值的正弦测试信号都能很方便地得到,所以很容易通过实验的方法简单可靠地得到系统的频率响应。我们从 1.4 节将会看到,一个未知的系统传递函数可以从系统的实验所确定的频率响应中导出。同时,在频域内进行系统设计,设计者能对系统带宽进行控制,还能对系统对噪声和干扰作出的响应进行衡量。

频率响应法的另一个优点是,把系统传递函数 $T(s)$ 中的 s 换成 $j\omega$,就可以得到描述系统正弦稳态响应的传递函数。这样描述系统正弦稳态响应的传递函数就是复变量 $j\omega$ 的函数,而其本身成为一个具有某个幅值和相角的复函数 $T(j\omega)$ 。 $T(j\omega)$ 的幅值和相角很容易用图形曲线来表示,从而能深入进行控制系统分析设计。

频率响应法在分析设计时的主要缺点是频域和时域之间的关联不直观。频率响应和对应的瞬态响应特征之间的直接对应关系是难以确定的。在实践中,通常是在各种各样的设计标准下,通过对频率响应进行调整,以获得令人满意的瞬态响应。

2.4 节给出拉普拉斯变换对,其形式为

$$F(s) = L\{f(t)\} = \int_0^{\infty} f(t)e^{-st} dt \quad (4.4.2)$$

及

$$f(t) = L^{-1}\{F(s)\} = \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} F(s)e^{st} ds \quad (4.4.3)$$

其中复变量 $s = \sigma + j\omega$ 。与此相似,傅里叶变换对可写为

$$F(j\omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{+\infty} f(t)e^{-j\omega t} dt \quad (4.4.4)$$

及

$$f(t) = \mathcal{F}^{-1}\{F(j\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(j\omega)e^{j\omega t} d\omega \quad (4.4.5)$$

对于 $f(t)$, 当

$$\int_{-\infty}^{+\infty} |f(t)| dt < \infty$$

时,傅里叶变换成立。

考查式(4.4.2)和式(4.4.4),可以看出,傅氏变换和拉氏变换有着密切的关系。当函数 $f(t)$ 仅定义在 $t \geq 0$ 时(这是一般情况),两者积分下限相同。那么我们可以注意到两式的不同仅在复变量之上。所以,如果函数 $f_1(t)$ 的拉氏变换为 $F_1(s)$,只要在 $F_1(s)$ 中令 $s = j\omega$,即可得到同一时间内的函数的傅里叶变换 $F_1(j\omega)$ 。

我们可能再次发问,既然傅氏变换和拉氏变换关系如此密切,为何我们不都使用拉氏变换而总是用傅里叶变换?从第 7 章可知,通过拉氏变换我们得以研究传递函数 $T(s)$ 的极点和零点在 s -平面上的分布。而通过频率响应法,我们研究的是传递函数 $T(j\omega)$ 及系统的幅值和相位特性。具备通过幅值、相位关系式及曲线图来表示和研究系统的特性的能力在进行控制系统分析设计时是很有用的。

如果我们考虑一个闭环系统的频率响应,我们可以有一个输入 $r(t)$,其频域内的傅氏变换如下:

$$R(j\omega) = \int_{-\infty}^{+\infty} r(t)e^{-j\omega t} dt$$

那么,单闭环控制系统的输出频率响应可以在闭环系统关系式 $Y(s) = T(s)R(s)$ 中通过 $s = j\omega$ 替换得到,所以我们有

$$Y(j\omega) = T(j\omega)R(j\omega) = \frac{G(j\omega)}{1 + G(j\omega)H(j\omega)}R(j\omega) \tag{4.4.6}$$

运用傅氏变换,输出瞬态响应应为

$$y(t) = \mathcal{F}^{-1}\{Y(j\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} Y(j\omega)e^{j\omega t} d\omega \tag{4.4.7}$$

然而,除了最简单的系统外,一般很难估算出这个反变换积分值,所以可以采用图形研究法。另外,在后续的章节中将会看到,瞬态响应的几个测量量都和频率特性有关,可用于设计之中。

频率响应图

系统的传递函数 $G(s)$ 可以在频域内用关系式

$$G(j\omega) = G(s) \big|_{s=j\omega} = R(\omega) + jX(\omega) \tag{4.4.8}$$

来表达,其中 $R(\omega) = \text{Re}[G(j\omega)]$, $X(\omega) = \text{Im}[G(j\omega)]$ 。(参阅 MCS 网站可以复习有关复数的内容。)

换一个方法,传递函数可以用形如

$$G(j\omega) = |G(j\omega)|e^{j\phi(j\omega)} = |G(j\omega)| \angle \phi(\omega) \tag{4.4.9}$$

的幅值 $|G(j\omega)|$ 和相位 $\phi(j\omega)$ 来表示。式中 $\phi(\omega) = \arctan \frac{X(\omega)}{R(\omega)}$, $|G(\omega)|^2 = [R(\omega)]^2 + [X(\omega)]^2$ 。

系统频率响应 $G(j\omega)$ 的图形表达可以用式(4.4.8)或式(4.4.9)来实现。频率响应的极图是用式(4.4.8)得到的。极图坐标由 $G(j\omega)$ 的实部和虚部构成,如图 4.4.1 所示。以下极图的例子可以说明如何使用这种方法。

例 4.4.1 RC 过滤器

图 4.4.2 所示为一个简单的 RC 过滤器,其传递函数为

$$G(s) = \frac{V_2(s)}{V_1(s)} = \frac{1}{RSc + 1} \tag{4.4.10}$$

正弦稳态传递函数为

$$G(j\omega) = \frac{1}{j\omega(RC) + 1} = \frac{1}{j(\omega/\omega_1) + 1} \tag{4.4.11}$$

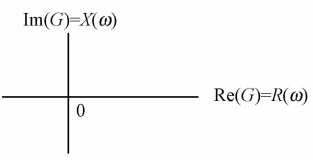


图 4.4.1

其中 $\omega_1 = \frac{1}{RC}$ 。则可从如下关系式得到极图：

$$\begin{aligned} G(j\omega) &= R(\omega) + jX(\omega) \\ &= \frac{1 - j(\omega/\omega_1)}{(\omega/\omega_1)^2 + 1} \\ &= \frac{1}{1 + (\omega/\omega_1)^2} - \frac{j(\omega/\omega_1)}{1 + (\omega/\omega_1)^2} \end{aligned} \tag{4.4.12}$$

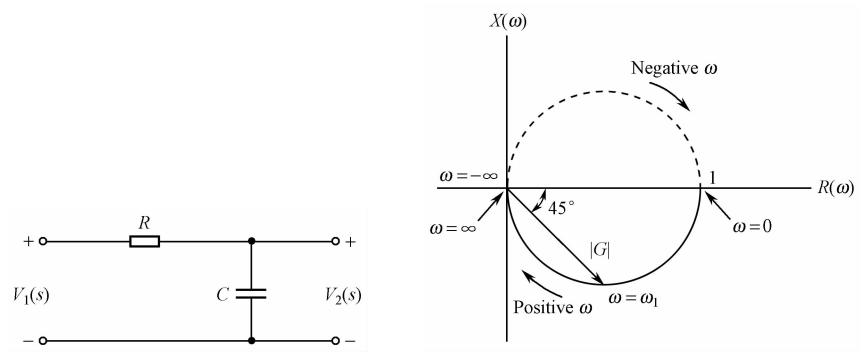


图 4.4.2 RC 滤波器

图 4.4.3 滤波器极坐标图

第一步是确定 $R(\omega)$ 和 $X(\omega)$ 在 $\omega = 0$ 和 $\omega = \infty$ 这两个频率下的值。 $\omega = 0$ 时可得 $R(\omega) = 1$ 和 $X(\omega) = 0$ 。 $\omega = \infty$ 时可得 $R(\omega) = 0$ 和 $Z(\omega) = 0$ 。这两个点在图 4.4.3 中示出。实部和虚部的根轨迹也示于图 4.4.3 中,而且容易看出它是一个以 $(1/2,0)$ 为圆心的圆。当 $\omega = \omega_1$,实部和虚部相等,而角 $\phi(\omega) = 45^\circ$ 。极图也可以直接得自式(4.4.9)：

$$G(j\omega) = |G(\omega)| + \angle \phi(\omega) \tag{4.4.13}$$

其中 $|G(\omega)| = \frac{1}{\sqrt{(\omega/\omega_1)^2 + 1}}^{1/2}$, $\phi(\omega) = -\arctan(\omega/\omega_1)$ 。

所以,当 $\omega = \omega_1$ 时,幅值为 $|G(\omega)| = \frac{1}{\sqrt{2}}$,相位 $\phi(\omega_1) = -45^\circ$ 。当 ω 趋向 $+\infty$ 时,有 $|G(\omega)| \rightarrow 0$ 及 $\phi(\omega_1) = -90^\circ$ 。与此相似,当 $\omega = 0$ 时,有 $|G(\omega)| = 1$, $\phi(\omega_1) = 0$ 。

4.4.5 Reading Materials

Frequency-Domain Approach

One of the first mathematical analysis of control systems was the frequency-domain approach. This is based on the developments of Pierre-Simon de Laplace (1749—1827), Joseph Fourier (1768—1830), Augustin Louis Cauchy (1789—1857), and others. The central concept of frequency-domain approach is that of transfer function. The transfer function of a linear time-invariant system is defined as $Y(s)/U(s)$, where $Y(s)$ is the Laplace transform of the output, and $U(s)$ is the Laplace transform of the input of the

system. It turns out that the transfer function is the Laplace transform of the system impulse response $h(t)$. Therefore, $H(s) = Y(s)/U(s)$, i. e. $H(s)$ embodies(具有) the transfer characteristics(传递特性) of the system. This approach is appropriate for linear time-invariant systems, especially for single-input/single-output systems where the graphical techniques are very efficient.

Frequency-domain approach originated in the process of solving of a major problem referring to the mass communication systems over long distances. To reduce distortions in amplifiers, after six years of persistence, Harold S. Black (1898-1983) revolutionized telecommunications by introducing the negative feedback in 1927(Black, 1934). As a method of system control, this has had a great impact in a large number of applications.

The theory of design the stable amplifiers was developed by Harry Nyquist (1889-1976) at Bell Laboratories. He derived his stability criterion, generally called the Nyquist stability theorem, based on the polar plot of the transfer function (Nyquist, 1932).

4.5 Routh's Stability Criterion

4.5.1 Text

The most important problem in linear control systems concerns stability. That is, under what conditions will a system become unstable? If it is unstable, how should we stabilize the system? Most linear closed-loop systems have closed-loop transfer functions of the form

$$\frac{C(s)}{R(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_{m-1} s + b_m}{a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n} = \frac{B(s)}{A(s)}$$

where the a 's and b 's are constants and $m \leq n$. A simple criterion, known as Routh's stability criterion, enables us to determine the number of closed-loop poles that lie in the right-half s plane without having to factor the denominator polynomial. (The polynomial may include parameters that MATLAB cannot handle.)

Routh's Stability Criterion. Routh's stability criterion tells us whether or not there are unstable roots in a polynomial equation without actually solving for them^[1]. This stability criterion applies to polynomials with only a finite number of terms. When the criterion is applied to a control system, information about absolute stability can obtained directly from the coefficients of the characteristic equation.

The procedure in Routh's stability criterion is as follows.

1. Write the polynomial in s in the following form:

$$a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n = 0 \tag{4.5.1}$$

where the coefficients are real quantities. We assume that $a_n \neq 0$; that is, any zero root has been removed.

2. If any of the coefficients are zero or negative in the presence of at least one positive coefficient, there is a root or roots that are imaginary or that have positive real parts^[2]. Therefore, in such a case, the system is not stable. If we are interested in only the absolute stability, there is no need to follow the procedure further. Note that all the coefficients must be positive. This is a necessary condition, as may be seen from the following argument: A polynomial in s having real coefficients can always be factored into linear and quadratic factors, such as $(s + a)$ and $(s^2 + bs + c)$, where a , b , and c are real. The linear factors yield real roots and the quadratic factors yield complex-conjugate roots of the polynomial. The factor $(s^2 + bs + c)$ yields roots having negative real parts only if b and c are both positive. For all roots to have negative real parts, the constants a , b , c , and so on, in all factors must be positive. The product of any number of linear and quadratic factors containing only positive coefficients always yields a polynomial with positive coefficients^[3]. It is important to note that the condition that all the coefficients be positive is not sufficient to assure stability. The necessary but not sufficient condition for stability is that the coefficients of Equation (4. 5. 1) all be present and all have a positive sign. (If all a 's are negative, they can be made positive by multiplying both sides of the equation by -1 .)

3. If all coefficients are positive, arrange the coefficients of the polynomial in rows and columns according to the following pattern:

$$\begin{array}{cccccc}
 s^n & a_0 & a_2 & a_4 & a_6 & \cdots \\
 s^{n-1} & a_1 & a_3 & a_5 & a_7 & \cdots \\
 s^{n-2} & b_1 & b_2 & b_3 & b_4 & \cdots \\
 s^{n-3} & c_1 & c_2 & c_3 & c_4 & \cdots \\
 s^{n-4} & d_1 & d_2 & d_3 & d_4 & \cdots \\
 \vdots & \vdots & \vdots & & & \\
 s^2 & e_1 & e_2 & & & \\
 s^1 & f_1 & & & & \\
 s^0 & g_1 & & & &
 \end{array}$$

The processes of forming rows continue until we run out of elements. (The total number of rows is $n+1$.) The coefficients b_1 , b_2 , b_3 , and so on, are evaluated as follows:

$$\begin{aligned}
 b_1 &= \frac{a_1 a_2 - a_0 a_3}{a_1} \\
 b_2 &= \frac{a_1 a_4 - a_0 a_5}{a_1} \\
 b_3 &= \frac{a_1 a_6 - a_0 a_7}{a_1}
 \end{aligned}$$

The evaluation of the b 's is continued until the remaining ones are all zero. The

same pattern of cross-multiplying the coefficients of the two previous rows is followed evaluating the c 's, d 's, e 's, and so on. That is,

$$\begin{aligned} c_1 &= \frac{b_1 a_3 - a_1 b_2}{b_1} \\ c_2 &= \frac{b_1 a_5 - a_1 b_3}{b_1} \\ c_3 &= \frac{b_1 a_7 - a_1 b_4}{b_1} \\ &\vdots \end{aligned}$$

and

$$\begin{aligned} d_1 &= \frac{c_1 b_2 - b_1 c_2}{c_1} \\ d_2 &= \frac{c_1 b_3 - b_1 c_3}{c_1} \\ &\vdots \end{aligned}$$

This process is continued until the n th row has been completed. The complete array of coefficients is triangular. Note that in developing the array an entire row may be divided or multiplied by a positive number in order to simplify the subsequent numerical calculation without altering the stability conclusion^[4].

Routh's stability criterion states that the number of roots of Equation (4.5.1) with positive real parts is equal to the number of changes in sign of the coefficients of the first column of the array. It should be noted that the exact values of the terms in the first column need not be known; instead, only the signs are needed. The necessary and sufficient condition that all roots of Equation (4.5.1) lie in the left-half s plane is that all the coefficients of Equation (4.5.1) be positive and all terms in the first column of the array have positive signs^[5].

Example 4.5.1

Let us apply Routh's stability criterion to the following third-order polynomial:

$$a_0 s^3 + a_1 s^2 + a_2 s + a_3 = 0$$

where all the coefficients are positive numbers. The array of coefficients becomes

$$\begin{array}{ccc} s^3 & a_0 & a_2 \\ s^2 & a_1 & a_3 \\ s^1 & \frac{a_1 a_2 - a_0 a_3}{a_1} & \\ s^0 & a_3 & \end{array}$$

The condition that all roots have negative real parts is given by

$$a_1 a_2 > a_0 a_3$$

Special Case. If a first-column term in any row is zero, but the remaining terms are not zero or is no remaining term, then the zero term is replaced by a very small positive number ϵ and the rest of the array is evaluated. For example, consider the following equation:

$$s^3 + 2s^2 + s + 2 = 0$$

The array of coefficients is

$$\begin{array}{rcl} s^3 & 1 & 1 \\ s^2 & 2 & 2 \\ s^1 & 0 \approx \epsilon & \\ s^0 & 2 & \end{array}$$

If the sign of the coefficient above the zero(ϵ) is the same as that below it, it indicates that there are a pair of imaginary roots.

If, however, the sign of the coefficient above the zero(ϵ) is opposite that below it, it indicates that there is one sign change. For example, for the equation

$$s^3 - 3s^2 + 2 = (s - 1)^2 (s + 2) = 0$$

the array of coefficients is

One sign change:

$$\begin{array}{rcl} s^3 & 1 & -3 \\ s^2 & 0 \approx \epsilon & 2 \end{array}$$

One sign change:

$$\begin{array}{rcl} s^1 & -3 - \frac{\epsilon}{2} & \\ s^0 & 2 & \end{array}$$

There are two sign changes of the coefficients in the first column. This agrees with the correct result indicated by the factored form of the polynomial equation.

If all the coefficients in any derived row are zero, it indicates that there are roots of equal magnitude lying radically opposite in the s plane, that is, two real roots with equal magnitudes and opposite signs and/or two conjugate imaginary roots. In such a case, the evaluation of the rest of the array can be continued by forming an auxiliary polynomial with the coefficients of the last row and by using the coefficients of the derivative of this polynomial in the next row. Such roots with equal magnitudes and lying radically opposite in the s plane can be found by solving the auxiliary polynomial, which is always even. For a $2n$ -degree auxiliary polynomial, there are n pairs of equal and opposite roots. For example, consider the following equation:

$$s^5 + 2s^4 + 24s^3 + 48s^2 - 25s - 50 = 0$$

The array of coefficients is

$$\begin{array}{cccc}
s^5 & 1 & 24 & -25 \\
s^4 & 2 & 48 & -50 \\
s^3 & 0 & 0 &
\end{array} \leftarrow \text{Auxiliary polynomial } P(s)$$

The terms in the s^3 row are all zero. (Note that such a case occurs only in an odd-numbered row.) The auxiliary polynomial is then formed the coefficients of the s^4 row. The auxiliary polynomial $P(s)$ is

$$P(s) = 2s^4 + 48s^2 - 50$$

Which indicates that there are two pairs of roots of equal magnitude and opposite sign (that is, two real roots with the same magnitude but opposite signs or two complex-conjugate roots on the imaginary axis). These pairs are obtained by solving the auxiliary polynomial equation $P(s) = 0$. The derivative of $P(s)$ with respect to s is

$$\frac{dP(s)}{ds} = 8s^3 + 96s$$

The terms in the s^3 row are replaced by the coefficients of the last equation, that is, 8 and 96. The array of coefficients then becomes

$$\begin{array}{cccc}
s^5 & 1 & 24 & -25 \\
s^4 & 2 & 48 & -50 \\
s^3 & 8 & 96 & \\
s^2 & 24 & -50 & \\
s^1 & 112.7 & 0 & \\
s^0 & -50 & &
\end{array} \leftarrow \text{Coefficients of } dP(s)/ds$$

We see that there is one change in the first column of the new array. Thus, the original equation has one root with a positive real part. By solving for roots of the auxiliary polynomial equation,

$$2s^4 + 48s^2 - 50 = 0$$

We obtain

$$s^2 = 1, s^2 = -25$$

or

$$s = \pm 1, s = \pm j5$$

These two pairs of roots of $P(s)$ are a part of the roots of the original equation. As a matter of fact, the original equation can be written in factored form as follows:

$$(s + 1)(s - 1)(s + js)(s - js)(s + 2) = 0$$

Clearly, the original equation has one root with a positive real part.

4.5.2 Specialized English Words

stability criterion	稳定判据	递函数	
closed-loop transfer function	闭环传	denominator polynomial	分母多项式

factor 因数,因子,因式
quadratic factors 二次因式
characteristic equation 特征方程
real quantity 实数
positive real part 正实部
complex-conjugate 复共轭的
triangular 三角形的
exact value 精确值,准确值
array 阵列,数组
third-order polynomial 三阶多项式

auxiliary polynomial 辅助多项式
odd-numbered row 奇数行
imaginary axis 虚轴
derivative 导出
row 行
column 列
necessary but not sufficient condition
必要但不充分条件
linear factor 一次因式

4.5.3 Notes

[1] Routh's stability criterion tells us whether or not there are unstable roots in a polynomial equation without actually solving for them. 句中“whether or not”引导定语从句,全句可译为“劳斯稳定性判据告诉我们,不用解方程就可以知道一个多项式方程是否含有不稳定的根。”

[2] If any of the coefficients are zero or negative in the presence of at least one positive coefficient, there is a root or roots that are imaginary or that have positive real parts. 此句的主句是“there is a root or roots”。“If”引导条件状语“any ...coefficient”,“that”引导宾语从句“are imaginary or that have positive real parts”修饰“a root or roots”。全句可译为“在有至少一个正系数的情况下,如果任何一个系数为零或者为负,方程就有一个或多个虚根或者有正实部的根。”

[3] The product of any number of linear and quadratic factors containing only positive coefficients always yields a polynomial with positive coefficients. 句中“containing”是现在分词在句中做定语,用来修饰“The product of any number of linear and quadratic factors”。注意句中“linear”意为“线性的”,即“正比的,一次的”。介词短语“with positive coefficients”修饰“a polynomial”。全句可译为“所有系数都为正的一次和二次因式的积,一定可以得到一个正系数的多项式。”

[4] Note that in developing the array an entire row may be divided or multiplied by a positive number in order to simplify the subsequent numerical calculation without altering the stability conclusion. 句中“in order to”引导目的状语从句“simplify ...conclusion”。全句可译为“注意,计算这个阵列时,为了简化后续数值的计算,一整行可以乘以或除以同一个正数,而不会改变有关系统的稳定性的结论。”

[5] The necessary and sufficient condition that all roots of Equation (4.5.1) lie in the left-half s plane is that all the coefficients of Equation (4.5.1) be positive and all terms in the first column of the array have positive signs. 句中第一个“that”引导定语从句,第二个“that”引导表语主语“all the ...signs”,此句是虚拟语气用“(should)+动词原

形”。全句可译为“方程(4.5.1)的所有根都位于 s 的左半平面的充分必要条件是,方程(4.5.1)的所有系数都是正数并且阵列第一列的所有项都为正。”

4.5.4 Reference Translation

劳斯稳定性判据

线性控制系统中最重要的是稳定性问题。也就是说,系统在何种情况下会变得不稳定? 如果系统是不稳定的,我们如何使系统稳定? 大多数线性闭环系统都有如下形式的闭环传递函数:

$$\frac{C(s)}{R(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_{m-1} s + b_m}{a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n} = \frac{B(s)}{A(s)}$$

式中,系数 a_0, a_1, \cdots, a_n 和 b_0, b_1, \cdots, b_m 都是常数,且 $m \leq n$ 。有一个简单的判据,称为劳斯稳定性判据,使我们不用考虑多项式的分母(或分母多项式)就能确定右半 s 平面中闭环极点的个数(多项式中可能包括 MATLAB 不能处理的参数)。

劳斯稳定性判据。 劳斯稳定性判据告诉我们,不用解方程就可以知道一个多项式方程是否含有不稳定的根。这个稳定判据只适用于有限项数的多项式。当劳斯判据应用于控制系统时,有关绝对稳定性的信息可以直接从特征方程的系数获得。

劳斯稳定判据的应用步骤如下。

1. 写出以下形式的 s 多项式:

$$a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n = 0 \tag{4.5.1}$$

式中,系数都是实数。假定 $a_n \neq 0$,已经去除了所有的零根。

2. 在有至少一个正系数的情况下,如果任何一个系数为零或者为负,方程就有一个或多个虚根或者有正实部的根。因此,这种情况下,系统是不稳定的。如果我们只对绝对稳定感兴趣,就不需要进行后面的步骤了。需要指出的是,所有的系数都必须为正的。这是一个必要条件,可以从下面的讨论中得出:一个有实系数的 s 多项式总可分解为像 $(s + a)$ 和 $(s^2 + bs + c)$ 这样的一次和二次的两种因式的乘积。式中, a, b, c 都为实数。从多项式的一次因式得到实根,二次因式得到共轭复根。当 b 和 c 都为正数时,因式 $(s^2 + bs + c)$ 得到负实部根。对于所有有负实部的根,所有因式里的常数 a, b, c 等都必须为正。所有系数都为正的一次和二次因式的积,一定可以得到一个正系数的多项式。需要强调的是,所有的系数都为正并不是系统稳定的充分条件。保证系统稳定的必要但不充分条件是方程(4.5.1)的所有系数不为零,并且都为正(如果所有 a 的系数都是负的,在方程的两边都乘以 -1 就可以得到正系数)。

3. 若所有的系数为正,按下面的模式将多项式系数行列进行排布:

$$\begin{array}{ccccccc}
s^n & a_0 & a_2 & a_4 & a_6 & \cdots \\
s^{n-1} & a_1 & a_3 & a_5 & a_7 & \cdots \\
s^{n-2} & b_1 & b_2 & b_3 & b_4 & \cdots \\
s^{n-3} & c_1 & c_2 & c_3 & c_4 & \cdots \\
s^{n-4} & d_1 & d_2 & d_3 & d_4 & \cdots \\
\vdots & \vdots & \vdots & & & \\
s^2 & e_1 & e_2 & & & \\
s^1 & f_1 & & & & \\
s^0 & g_1 & & & &
\end{array}$$

该过程一直进行到所有系数处理完毕(总行数为 $n+1$)。系数 b_1, b_2, b_3 等按如下方式计算:

$$\begin{aligned}
b_1 &= \frac{a_1 a_2 - a_0 a_3}{a_1} \\
b_2 &= \frac{a_1 a_4 - a_0 a_5}{a_1} \\
b_3 &= \frac{a_1 a_6 - a_0 a_7}{a_1}
\end{aligned}$$

当剩下的数都为零时, b 系数的计算便结束。按照上述将前两行的系数交叉相乘的相同模式, 计算 c, d, e 等系数, 即

$$\begin{aligned}
c_1 &= \frac{b_1 a_3 - a_1 b_2}{b_1} \\
c_2 &= \frac{b_1 a_5 - a_1 b_3}{b_1} \\
c_3 &= \frac{b_1 a_7 - a_1 b_4}{b_1} \\
&\vdots \\
d_1 &= \frac{c_1 b_2 - b_1 c_2}{c_1} \\
d_2 &= \frac{c_1 b_3 - b_1 c_3}{c_1} \\
&\vdots
\end{aligned}$$

及

这种过程一直进行到第 n 行为止。最后, 系数阵列排列成三角形。注意, 计算这个阵列时, 为了简化后续数值的计算, 一整行可以乘以或除以同一个正数, 而不会改变有关系统的稳定性的结论。

劳斯判据表明, 方程(4.5.1)的正实部根的个数等于阵列第一列系数的符号变化次数。应该注意的是, 不需要知道第一列各项的精确值, 而只需要知道它们的符号。方程(4.5.1)的所有根都位于 s 的左半平面的充分必要条件是, 方程(4.5.1)的所有系数都是正数并且阵列第一列的所有项都为正。

例 4.5.1 应用劳斯判据判定以下三阶多项式的稳定性：

$$a_0 s^3 + a_1 s^2 + a_2 s + a_3 = 0$$

式中所有的系数都是正数。系数阵列为

$$\begin{array}{ccc} s^3 & a_0 & a_2 \\ s^2 & a_1 & a_3 \\ s^1 & \frac{a_1 a_2 - a_0 a_3}{a_1} & \\ s^0 & a_3 & \end{array}$$

所有的根为负实部的条件由下式给出：

$$a_1 a_2 > a_0 a_3$$

特殊情况。如果某行的第一列的项为零，而其余各项不为零或者没有其余项，那么为零的项可以用一个非常小的正数 ϵ 代替，则阵列的其他部分就可以继续计算了。例如，判断下面的方程：

$$s^3 + 2s^2 + s + 2 = 0$$

系数阵列为

$$\begin{array}{ccc} s^3 & 1 & 1 \\ s^2 & 2 & 2 \\ s^1 & 0 \approx \epsilon & \\ s^0 & 2 & \end{array}$$

如果零项(ϵ)上面系数的符号和下面系数的符号相同，就表明方程有一对虚根。

但是，如果零项(ϵ)上面系数的符号和下面系数的符号相反，就表明有一次符号变化。例如，方程

$$s^3 - 3s^2 + 2 = (s - 1)^2 (s + 2) = 0$$

的系数阵列为

一次符号变化：

$$\begin{array}{ccc} s^3 & 1 & -3 \\ s^2 & 0 \approx \epsilon & 2 \end{array}$$

一次符号变化：

$$\begin{array}{ccc} s^1 & -3 - \frac{\epsilon}{2} & \\ s^0 & 2 & \end{array}$$

系数阵列的第一列有两次符号变化。这与由多项式方程的因式分解形式所表明的正确结论是一致的。

如果某一导出行(即从第三行开始的行)的系数都为零，表明特征方程有一对大小相等的根完全对称地处在 s 平面上，即两个大小相等、符号相反的实根和/或两个共轭虚根。这种情况下，可用全零行的上一行的系数构造一个辅助方程，多项式的下一行的系数可以从这个辅助方程得出，这样，稳定判据就可以进行下去(阵列的余下部分可以继续推演如

下:用最后一行的系数构造一个辅助多项式,再下一行的系数又由对这一行求导数得出)。s 平面上像这样大小相等、符号相反的根都可以由辅助方程求得,且总是偶数。对于一个 $2n$ 阶辅助多项式,有 n 对大小相等、符号相反的根。例如,考虑下面的方程:

$$s^5 + 2s^4 + 24s^3 + 48s^2 - 25s - 50 = 0$$

系数阵列为

$$\begin{array}{cccc} s^5 & 1 & 24 & -25 \\ s^4 & 2 & 48 & -50 \\ s^3 & 0 & 0 & \end{array} \leftarrow \text{辅助多项式 } P(s)$$

s^3 这一行的项都为零(通常这种情况只发生在奇数行)。那么 s^4 行的系数由辅助多项式产生。辅助多项式 $P(s)$ 为

$$P(s) = 2s^4 + 48s^2 - 50$$

此式表明,辅助方程有两对数值相等、符号相反的根(即两个大小相等、符号相反的实根或者两个位于虚轴上的共轭复根)。这些成对的根是通过求解多项式方程 $P(s) = 0$ 得到的。 $P(s)$ 对 s 的导数为

$$\frac{dP(s)}{ds} = 8s^3 + 96s$$

s^3 行的项由上面方程的系数来替换,即由 8 和 96 替换。系数阵列变为

$$\begin{array}{cccc} s^5 & 1 & 24 & -25 \\ s^4 & 2 & 48 & -50 \\ s^3 & 8 & 96 & \\ s^2 & 24 & -50 & \\ s^1 & 112.7 & 0 & \\ s^0 & -50 & & \end{array} \leftarrow dP(s)/ds \text{ 的系数}$$

我们看到新阵列的第一列有一次符号变化。因此,原方程有一个正实根。求解辅助多项式方程

$$2s^4 + 48s^2 - 50 = 0$$

得到

$$s^2 = 1, s^2 = -25$$

或

$$s = \pm 1, s = \pm j5$$

$P(s)$ 的两对根是原方程根的一部分。实际上,原方程可以写成如下因式相乘的形式:

$$(s + 1)(s - 1)(s + js)(s - js)(s + 2) = 0$$

显然,原方程有一个正实根。

4.5.5 Reading Materials

Routh-Hurwitz Stability Criterion

The Routh-Hurwitz stability criterion is a necessary (and frequently sufficient) method to establish the stability of a single-input, single-output (SISO), linear time invariant (LTI) (线性时不变) control system. More generally, given a polynomial, some calculations using only the coefficients of that polynomial can lead to the conclusion that it is not stable. For the discrete case, see the Jury test equivalent.

The criterion establishes a systematic way to show that the linearized equations of motion of a system have only stable solutions $\exp(pt)$, that is where all p have negative real parts. It can be performed using either polynomial divisions or determinant calculus.

The criterion is derived through the use of the Euclidean algorithm and Sturm's theorem in evaluating Cauchy indices.

The criterion is related to Routh-Hurwitz theorem. Indeed, from the statement of that theorem, we have $p - q = \omega(+\infty) - \omega(-\infty)$.

where:

p is the number of roots of the polynomial $f(z)$ located in the left half-plane;

q is the number of roots of the polynomial $f(z)$ located in the right half-plane (let us remind ourselves that f is supposed to have no roots lying on the imaginary line);

$\omega(x)$ is the number of variations of the generalized Sturm chain obtained from $P_0(y)$ and $P_1(y)$ (by successive Euclidean divisions) where $f(iy) = P_0(y) + iP_1(y)$ for a real y .

By the fundamental theorem of algebra, each polynomial of degree n must have n roots in the complex plane (i. e. , for an f with no roots on the imaginary line, $p+q = n$). Thus, we have the condition that f is a (Hurwitz) stable polynomial(稳定多项式) if and only if $p-q = n$ (the proof is given below). Using the Routh-Hurwitz theorem, we can replace the condition on p and q by a condition on the generalized Sturm chain, which will give in turn a condition on the coefficients of f .

4.6 State Variable Methods

4.6.1 Text

Preview

In the preceding chapter, we developed and studied several useful approaches to the analysis and design of feedback systems. The Laplace transform was utilized to transform the differential equations representing the system to an algebraic equation

expressed in terms of the complex variable s . Utilizing this algebraic equation, we were able to obtain a transfer function representation of the input-output relationship.

With the ready availability of digital computers, it is convenient to consider the time-domain formulation of the equations representing control systems. The time-domain techniques can be utilized for nonlinear, time-varying, and multivariable systems.

A time-varying control system is a system for which one or more of the parameters of the system may vary as a function of time.

For example, the mass of a missile varies as a function of time as the fuel is expended during flight. A multivariable system, as discussed in Section 2.6, is a system with several input and output signals. The solution of a time-domain formulation of a control system problem is facilitated by the availability and ease of use of digital computers. Therefore we are interested in reconsidering the time-domain description of dynamic systems as they are represented by the system differential equation. The time-domain is the mathematical domain that incorporates the response and description of a system in terms of time, t .

The time-domain representation of control systems is an essential basis for modern control theory and system optimization. In Chapter 11, we will have an opportunity to design an optimum control system by utilizing time-domain methods.

The State Variables of a Dynamic

The time-domain analysis and design of control systems utilizes the concept of the state of a system.

The state of a system is a set of variables such that the knowledge of these variables and the input functions will, with the equations describing the innimagate future state and output of the system.

For a dynamic system, the state of a system described in terms of a set of state variables $[x_1(t), x_2(t), \cdots, x_n(t)]$. The state variables are those variables that

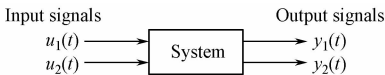


Figure 4. 6. 1 System block diagram.

determine the future behavior of a system when the present state of the system and the excitation signals are known. Consider the system shown in Figure 4. 6. 1, where $y_1(t)$ and $y_2(t)$ are the output signals and $u_1(t)$ and $u_2(t)$ are the input signals. A set of state variables (x_1, x_2, \cdots, x_n) for the system shown in the figure is a set such that knowledge of the initial values of the state variables $[x_1(t_0), x_2(t_0), \cdots, x_n(t_0)]$ at the initial time to, and of the input signals $u_1(t)$ and $u_2(t)$ for $t \geq t_0$, suffices to determine the future values of the outputs and state variables^[2].

The state variables describe the future response of a system, given the present state, the excitation inputs, and the equations describing the dynamics.

The general form of a dynamic system is shown in Figure 4. 6. 2.

A simple example of a state variable is the state of an on-off light switch. The switch can be in either the on or the off position, and thus the state of the switch can assume one of two possible values. Thus, if we know the present state (position) of the switch at t_0 and if an input is applied, we are able to determine the future value of the state of the element.

The concept of a net of state variables that represent a dynamic system can be illustrated in terms of the spring-mass-damper system shown in Figure 4. 6. 3. The number of state variables chosen to represent

this system should be as small as possible in order to avoid redundant state variable, A set of state variables sufficient to describe this system includes the position and the velocity of the mass. Therefore, we will define a set of state variables as (x_1, x_2) , where

$$x_1(t) = y(t) \quad \text{and} \quad x_2(t) = \frac{dy(t)}{dt}$$

The differential equation describes the behavior of the system and is usually written as

$$M \frac{d^2 y}{dt^2} + b \frac{dy}{dt} + ky = u(t) \tag{4. 6. 1}$$

To write Equation (4. 6. 1) in terms of the state variable, we substitute the state variables as already defined and obtain

$$M \frac{dx_2}{dt} + bx_2 + kx_1 = u(t) \tag{4. 6. 2}$$

Therefore, we can write the equations that describe the behavior of the spring-mass-damper system as the set of two first-order differential equations

$$\frac{dx_1}{dt} = x_2 \tag{4. 6. 3}$$

and

$$\frac{dx_2}{dt} = -\frac{b}{M}x_2 - \frac{k}{M}x_1 + \frac{1}{M}u \tag{4. 6. 4}$$

This set of differential equations describes the behavior of the state of the system in Figure 4. 6. 4.

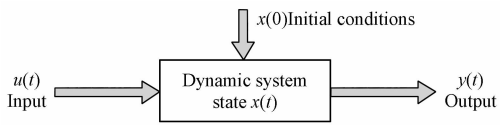


Figure 4. 6. 2 Dynamic system.

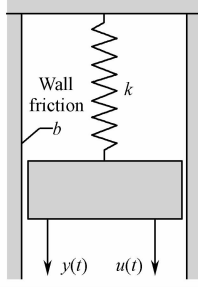


Figure 4.6.3 A spring-mass-damper system.

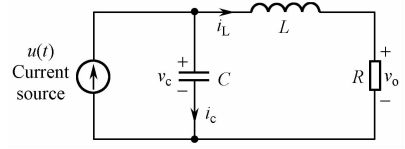


Figure 4.6.4 A RLC circuit.

As another example of the state variable characterization of a system, let us consider the RLC circuit shown in Figure 4.6.4. The state of this system can be described in terms of a set of state variables (x_1, x_2) where x_1 is the capacitor voltage $u_c(t)$ and x_2 is equal to the inductor current $i_L(t)$. This choice of state variables as

$$L = (1/2)Li_L^2 + (1/2)Cv_c^2 \quad (4.6.5)$$

Therefore $x_1(t_0)$ and $x_2(t_0)$ represent the total initial energy of the network and thus the state of the system at $t = t_0$. For a passive RLC network, the number of state variables required is equal to the number of independent energy-storage elements. Utilizing Kirchhoff's current law at the junction, we obtain a first-order differential equation by describing the rate of change of capacitor voltage as

$$i_c = C \frac{dv_c}{dt} = +u(t) - i_L \quad (4.6.6)$$

Kirchhoff's voltage law for the right-hand loop provides the equation describing the rate of change of inductor current as

$$L \frac{di_L}{dt} = -Ri_L + v_c \quad (4.6.7)$$

The output of this system is represented by the linear algebraic equation

$$v_o = Ri_L(t)$$

We can rewrite Equations (4.6.6) and (4.6.7) as a set of two first-order differential equations in terms of the state variables x_1 and x_2 as follows:

$$\frac{dx_1}{dt} = -\frac{1}{C}x_2 + \frac{1}{C}u(t) \quad (4.6.8)$$

and

$$\frac{dx_2}{dt} = +\frac{1}{L}x_1 - \frac{R}{L}x_2 \quad (4.6.9)$$

The output signal is then

$$y_1(t) = v_o(t) = Rx_2 \quad (4.6.10)$$

Utilizing Equations (4.6.8) and (4.6.9) and the initial conditions of the network

represented by $[x_1(t_0), x_2(t_0)]$, we can determine the system's future behavior and its output.

The State Differential Equation

The state of a system is described by the set of first-order differential equations written in terms of the state variables (x_1, x_2, \dots, x_n) . These first-order differential equations can be written in general form as

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + b_{11}u_1 + b_{12}u_2 + \dots + b_{1m}u_m \quad (4.6.11)$$

$$\dot{x}_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n + b_{21}u_1 + b_{22}u_2 + \dots + b_{2m}u_m \quad (4.6.12)$$

...

$$\dot{x}_n = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n + b_{n1}u_1 + b_{n2}u_2 + \dots + b_{nm}u_m \quad (4.6.13)$$

where $\dot{x} = dx/dt$. Thus, this set of simultaneous differential equations can be written in matrix form as follows:

$$\frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nm} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \quad (4.6.14)$$

The column matrix consisting of the state variables is called the state vector and is written as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (4.6.15)$$

where the boldface indicates a vector. The vector of input signals is defined as \mathbf{u} . Then the system can be represented by the compact notation of the state differential equation as

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad (4.6.16)$$

The differential equation (4.6.16) is also commonly called the state equation.

The matrix \mathbf{A} is an $n \times n$ square matrix, and \mathbf{B} is an $n \times m$ matrix. The state differential equation relates the rate of change of the state of the system to the state of the system and the input signals. In general, the outputs of a linear system can be related to the state variables and the input signals by the output equation

$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du} \quad (4.6.17)$$

where \mathbf{y} is the set of output signals expressed in column vector form. The state-space representation (or state-variable representation) is comprised of the state differential equation and the output equation.

We use Equations (4.6.8) and (4.6.9) to obtain the state variable differential

equation for the RLC of Figure 4. 6. 4 as

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -1/C \\ 1/L & -R/L \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1/C \\ 0 \end{pmatrix} \mathbf{u}(t) \quad (4. 6. 18)$$

and the output as

$$\mathbf{y} = [0 \quad R] \mathbf{x} \quad (4. 6. 19)$$

When $R = 3$, $L = 1$, and $C = 1/2$, we have

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -2 \\ 1 & -3 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} \mathbf{u}$$

and

$$\mathbf{y} = [0 \quad 3] \mathbf{x}$$

The solution of the state differential equation (Equation (4. 6. 16)) can be obtained in a manner similar to the approach we utilize for solving a first-order differential equation.

4. 6. 2 Specialized English Words

algebraic	代数的	column	列(矩阵)
time-domain	时域	state vector	状态矢量
nonlinear, time-varying, and systems	非线性时变系统	boldface = bold type	黑体, 粗体
formulation	制定, 规划, 确切表达	square matrix	矩形矩阵, 方阵
suffice	足够, 足以	capacitor	电容
redundant	冗余的	inductor	电感
matrix	矩阵	intuitively	(凭)直觉
		passive	无源的, 被动的

4. 6. 3 Notes

[1]For example, the mass of a missile varies as a function of time as the fuel is expended during flight. 这是一个主从复合句。主句为“the mass of a missile varies”。注意后面的“as...as...”并非那个常见的“as ...as...”(像……一样 ……)短语,而是分别由“as”引起的两个状语:前一个“as”是介词,“as a function of time”是介词短语,做状语,说明“varies”的方式和特点;后一个“as”是连词,引起一状语从句,表明“varies”的条件。全句可译为“例如,火箭的质量随着燃料在飞行中的消耗而成为随时间变化的函数。”

[2]The state of a system is a set of variables such that ... 在这个复合句中,“that”引起的定语从句修饰“a set of variables”,注意“such”的用法,它代指后面从句的内容。“a set of variables such that ...”等于“such a set of variables that ...”,这样语法关系显得更清楚。全句可译为“系统状态是这样一组变量,有了这组变量的信息加上输入函数,就能通过描述系统动态过程的方程,给出系统的未来状态和输出。”

4.6.4 Reference Translation

状态变量模型

引言

在前一章中,我们推演分析了几个十分有用的反馈系统的分析和设计方法。通过拉氏变换将描述系统的微分方程转换为用复变量 s 表示的代数方程,再运用这个代数方程,就能得到表示输入/输出关系的传递函数表达式。

借助于数字计算机现有的计算能力,可以很方便地建立代表系统的诸等式的时域表达。时域技术可用于非线性时变多变量系统。

时变控制系统指的是系统中的一个或多个参数是随时间变化的函数。

例如,火箭的质量随着燃料在飞行中的消耗而成为随时间变化的函数。如 2.6 节所讨论的,多变量系统是具有多个输入输出信号的系统。由于数字计算机的普及便捷,控制系统问题的时域方程求解是很容易的。所以我们有兴趣重新考虑用系统微分方程表示的动态系统的时域表达。时域是用时间 t 对系统及其响应进行表述的数学域。

控制系统的时域表达是现代控制理论和系统优化的根本性基础。在第 11 章中,我们将有机会用时域方法去设计一个系统的最优控制。在本章中,我们将推导控制系统时域表达式,并用图解说明求解系统时间响应的几种方法。

动态系统的状态变量

控制系统的时域分析和设计采用了系统状态的概念。

系统状态是这样一组变量,有了这组变量的信息加上输入函数,就能通过描述系统动态过程的方程,给出系统的未来状态和输出。

对于一个动态系统,系统状态是用一组状态变量 $[x_1(t), x_2(t), \dots, x_n(t)]$ 来描述的。状态变量是已知系统现在的状态以及激励信号,就能决定系统将来的行为的那些变量。分析图 4.6.1

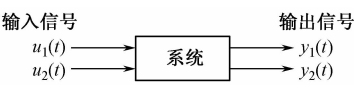


图 4.6.1 系统框图

中的系统,其中 $y_1(t)$ 和 $y_2(t)$ 为输出信号, $u_1(t)$ 和 $u_2(t)$ 为输入信号。图中的系统的一组状态变量,有了初始时刻 t_0 时的状态变量的初始值 $[x_1(t_0), x_2(t_0), \dots, x_n(t_0)]$, 以及 $t \geq t_0$ 时的输入信号 $u_1(t)$ 和 $u_2(t)$ 的初始值,就足以确定输出信号及状态变量的将来值。

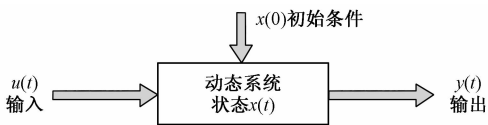


图 4.6.2 动态系统

图 4.6.2 给出了动态系统的一般形式。

给定现在的状态、激励输入以及描写系统动态特性的方程,状态变量就可以描述系统的未来响应。

灯具的通断开关就是状态变量的简单例子。开关可以处于通或断的位置上,所以开关的状态可以假定为两种可能之一。如果我们知道开关在 t_0 时刻现在的状态(位置),而且有了电源输入,我们就能确定灯具状态的未来之值。

用一组状态变量来表示动态系统的概念可以用图 4.6.3 所示的弹簧-质量-阻尼器系统予以说明。所选用来表示这个系统的状态变量的数量应当尽量少,避免出现冗余的状态变量。一组足以描述这个系统的状态变量包括质量的位置和速度。这样,我们可以定义一组状态变量 (x_1, x_2) , 其中,

$$x_1(t) = y(t), x_2(t) = \frac{dy(t)}{dt}$$

微分方程可用来描述这个系统的行为,一般记为

$$M \frac{d^2 y}{dt^2} + b \frac{dy}{dt} + ky = u(t) \tag{4.6.1}$$

为了用状态变量建立方程(4.6.1),我们用已经定义的状态变量进行替换,可得

$$M \frac{dx_2}{dt} + bx_2 + kx_1 = u(t) \tag{4.6.2}$$

所以,我们可以将描述弹簧-质量-阻尼器系统行为的方程确定为由两个一阶微分方程

$$\frac{dx_1}{dt} = x_2 \tag{4.6.3}$$

及

$$\frac{dx_2}{dt} = -\frac{b}{M}x_2 - \frac{k}{M}x_1 + \frac{1}{M}u \tag{4.6.4}$$

所组成的方程组。这个方程组通过每个状态变量的变化率来描述系统状态的行为。

让我们把图 4.6.4 所示的 RLC 电路作为状态变量特征表达的另一个例子。该系统的状态可以用一组状态变量 (x_1, x_2) 来描述,这里 x_1 是电容电压 $u_c(t)$, x_2 是电感电流 $i_L(t)$ 。这样选择状态变量显然是恰当的,因为电路(网络)所存储的能量可以用这些变量表示为

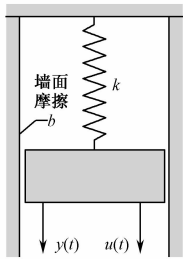


图 4.6.3 弹簧-质量-阻尼器系统

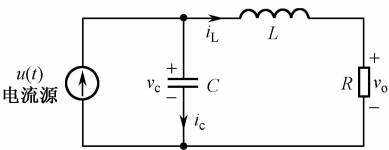


图 4.6.4 RLC 电路

$$L = (1/2)Li_L^2 + (1/2)Cv_c^2 \tag{4.6.5}$$

所以 $x_1(t_0)$ 和 $x_2(t_0)$ 代表电路的初始总能量,即 $t = t_0$ 时的状态。对于无源 RLC 电路,所需的状态变量数量等于独立储能元件的数量。对节点运用基尔霍夫电流定律描述电容电压变化率,我们可以得到一个一阶微分方程:

$$i_c = C \frac{dv_c}{dt} = +u(t) - i_L \quad (4.6.6)$$

对图中右边的环路运用基尔霍夫电压定律,可以得到描述电感电流变化率的方程:

$$L \frac{di_L}{dt} = -Ri_L + v_c \quad (4.6.7)$$

该系统输出是由线性代数方程

$$v_o = Ri_L(t)$$

来描述的。

通过状态变量 x_1 和 x_2 , 我们可以将式(4.6.6)和式(4.6.7)改写为一个由两个一阶微分方程组成的方程组:

$$\frac{dx_1}{dt} = -\frac{1}{C}x_2 + \frac{1}{C}u(t) \quad (4.6.8)$$

及

$$\frac{dx_2}{dt} = +\frac{1}{L}x_1 - \frac{R}{L}x_2 \quad (4.6.9)$$

输出信号则为

$$y_1(t) = v_o(t) = Rx_2 \quad (4.6.10)$$

通过式(4.6.8)和式(4.6.9)及由 $[x_1(t_0), x_2(t_0)]$ 表示的电路初始条件,我们可以确定电路的未来行为和输出。

状态微分方程

系统的状态可以用由状态变量 (x_1, x_2, \dots, x_n) 写成的一阶微分方程组来描述。这些一阶微分方程的一般形式为

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + b_{11}u_1 + b_{12}u_2 + \dots + b_{1m}u_m \quad (4.6.11)$$

$$\dot{x}_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n + b_{21}u_1 + b_{22}u_2 + \dots + b_{2m}u_m \quad (4.6.12)$$

...

$$\dot{x}_n = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n + b_{n1}u_1 + b_{n2}u_2 + \dots + b_{nm}u_m \quad (4.6.13)$$

其中 $\dot{x} = dx/dt$ 。所以,这一组同步微分方程组可以用矩阵的形式表示如下:

$$\frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nm} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \quad (4.6.14)$$

由状态变量组成的行矩阵称为状态矢量,记为

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (4.6.15)$$

其中粗体字表示矢量。输入信号矢量定义为 \mathbf{u} 。这样,系统就可以用状态微分方程的省写形式表示为

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (4.6.16)$$

微分方程(4.6.16)通常也称为状态方程。

矩阵 \mathbf{A} 是一个 $n \times n$ 的方阵, \mathbf{B} 是一个 $n \times m$ 矩阵。状态微分方程把系统状态的变化率和系统的状态以及输入信号联系了起来。一般而言,一个线性系统的输出通过输出方程

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad (4.6.17)$$

将它和状态变量以及输入信号联系起来。方程中 \mathbf{y} 是用行矩阵形式表示的一组输出信号。状态空间表达式(或状态变量表达式)由状态微分方程和输出方程组成。

我们用式(4.6.8)和式(4.6.9)得到图 4.6.4 所示 RLC 电路的状态空间微分方程

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -1/C \\ 1/L & -R/L \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1/C \\ 0 \end{pmatrix} \mathbf{u}(t) \quad (4.6.18)$$

及输出

$$\mathbf{y} = [0 \quad R]\mathbf{x} \quad (4.6.19)$$

当 $R = 3, L = 1, C = 1/2$ 时,我们有

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -2 \\ 1 & -3 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} \mathbf{u}$$

及

$$\mathbf{y} = [0 \quad 3]\mathbf{x}$$

状态微分方程[式(4.6.16)]的解可以用与求解一阶微分方程的类似方法求得。

4.6.5 Reading Materials

Time-Domain Algebraic Approach

The modern era in control theory started with the work of Rudolf Kalman who published a number of books in which the main problems of nonlinear systems theory was presented. In [Kalman and Bertram, 1960] the Lyapunov stability in time-domain of nonlinear systems is considered. The optimal control of systems as well as the design of linear quadratic regulator(线性二阶型调节器) is discussed in [Kalman, 1960a]. The optimal filtering and estimation theory, and the design equation for the discrete Kalman filter (卡尔曼滤波) was presented in [Kalman, 1960b]. The continuous version of

Kalman filter was developed in [Kalman and Bucy, 1961]. To overcome the limitation of frequency-domain approach, which is very much an art and provide a non-unique feedback, Kalman introduced the concept of “state”, a mathematical entity(数学概念) that mediates(处于两者之间, 居中) between inputs and outputs. The importance of this concept is based on the fact that the state of a dynamical system emphasizes the notions(理解, 认识) of causality(因果关系) and internal structure. For finite-dimensional systems, i. e. systems for which the state belongs to a finite-dimensional vector space, the representation is given by a first-order vector differential equation of the following form:

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t)\end{aligned}$$

where \mathbf{x} is the vector of internal variables, or system states, $\mathbf{u}(t)$ is the vector of control inputs and $\mathbf{y}(t)$ is the vector of measured outputs. The matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} describes the system dynamical interconnections(相互关系).

4.7 Root-Locus

4.7.1 Text

The basic characteristic of the transient response of a closed-loop system is closely related to the location of the closed-loop poles. If the system has a variable loop gain, then the location of the closed-loop poles depends on the value of the loop gain chosen. It is important, therefore, that the designer knows how the closed-loop poles move in the s plane as the loop gain is varied^[1].

From the design viewpoint, in some systems simple gain adjustment may move the closed-loop poles to desired locations. Then the design problem may become the selection of an appropriate gain value. If the gain adjustment alone does not yield a desired result, addition of a compensator to the system will become necessary.

The closed-loop poles are the roots of the characteristic equation. Finding the roots of the characteristic equation of degree higher than 3 is laborious and will need computer solution. (MATLAB provides a simple solution to this problem.) However just finding the roots of the characteristic equation may be of limited value, because as the gain of the open-loop transfer function varies the characteristic equation changes and the computations must be repeated.

A simple method for finding the roots of the characteristic equation has been developed by W. R. Evans and used extensively in control engineering. This method, called the root-locus method, is one in which the roots of the characteristic equation are plotted for all values of a system parameter. The roots corresponding to a particular

value of this parameter can then be located on the resulting graph. Note that the parameter is usually the gain, but any other variable of the open-loop transfer function may be used. Unless otherwise stated, we shall assume that the gain of the open-loop transfer function is the parameter to be varied through all values, from zero to infinity.

By using the root-locus method the designer can predict the effects on the location of the closed-loop poles of varying the gain value or adding open-loop poles and/or open-loop zeros. Therefore, it is desired that the designer have a good understanding of the method for generating the root locus of the closed-loop system, both by hand and by use of a computer software like MATLAB.

Root-Locus Method. The basic idea behind the root-locus method is that the values of s that make the transfer function around the loop equal -1 must satisfy the characteristic equation of the system.

The root locus is the locus of roots of the characteristic equation of the closed-loop system as a specific parameter (usually, gain K) is varied from zero to infinity, giving the method its name^[2]. Such a plot clearly shows the contributions of each open-loop pole or zero to the locations of the closed-loop poles.

In designing a linear control system, we find that the root-locus method proves quite useful since it indicates the manner in which the open-loop poles and zeros should be modified so that the response meets system performance specifications^[3]. This method is particularly suited to obtaining approximate results very quickly.

Because generating the root loci by use of MATLAB is very simple, one may think sketching the root loci by hand is a waste of time and effort. However, experience in sketching the root loci by hand is invaluable for interpreting computer-generated root loci, as well as for getting a rough idea of the root loci very quickly.

By using the root-locus method, it is possible to determine the value of the loop gain K that will make the damping ratio of the dominant closed-loop poles as prescribed. If the location of an open-loop pole or zero is a system variable, then the root-locus method suggests the way to choose the location of an open-loop pole or zero.

Angle and Magnitude Conditions. Consider the system shown in Figure 4.7.1. The closed-loop transfer function is

$$\frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} \quad (4.7.1)$$

The characteristic equation for this closed-loop system is obtained by setting the denominator of the right-hand side of Equation (4.7.1) equal to zero. That is,

$$1 + G(s)H(s) = 0$$

or

$$G(s)H(s) = -1 \quad (4.7.2)$$

Here we assume that $G(s)H(s)$ is a ratio of polynomials in s . Since $G(s)H(s)$ is a complex quantity, Equation (4.7.2) can be split into two equations by equating the angles and magnitudes of both sides, respectively, to obtain the following:

Angle condition:

$$\angle G(s)H(s) = \pm 180^\circ(2k+1), k = 0, 1, 2, \dots \quad (4.7.3)$$

Magnitude condition:

$$|G(s)H(s)| = 1 \quad (4.7.4)$$

The values of s that fulfill both the angle and magnitude conditions are the roots of the characteristic equation, or the closed-loop poles. A locus of the points in the complex plane satisfying the angle condition alone is the root locus. The roots of the characteristic equation (the closed-loop poles) corresponding to a given value of the gain can be determined from the magnitude condition.

In many cases, $G(s)H(s)$ involves a gain parameter k , and the characteristic equation may be written as

$$1 + \frac{k(s+z_1)(s+z_2)\cdots(s+z_m)}{(s+p_1)(s+p_2)\cdots(s+p_n)} = 0$$

Then the root loci for the system are the loci of the closed-loop poles as the gain K is varied from zero to infinity^[4].

Note that to begin sketching the root loci of a system by the root-locus method we must know the location of the poles and zeros of $G(s)H(s)$. Remember that the angles of the complex quantities originating from the open-loop poles and open-loop zeros to the test point s are measured in the counterclockwise direction^[5]. For example, if $G(s)H(s)$ is given by

$$G(s)H(s) = \frac{k(s+z_1)}{(s+p_1)(s+p_2)(s+p_3)(s+p_4)}$$

where $-p_2$ and $-p_3$ are complex-conjugate poles, then the angle of $G(s)H(s)$ is

$$\angle G(s)H(s) = \phi_1 - \theta_1 - \theta_2 - \theta_3 - \theta_4$$

where $\phi_1, \theta_1, \theta_2, \theta_3$ and θ_4 are measured counterclockwise as shown in Figures 4.7.2(a) and (b). The magnitude of $G(s)H(s)$ for this system is

$$|G(s)H(s)| = \frac{kB_1}{A_1A_2A_3A_4}$$

where A_1, A_2, A_3, A_4 , and B_1 are the magnitudes of the complex quantities $s+p_1, s+p_2, s+p_3, s+p_4$, and $s+z_1$, respectively, as shown in Figure 4.7.2(a).

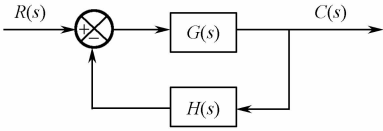


Figure 4.7.1 Control system.

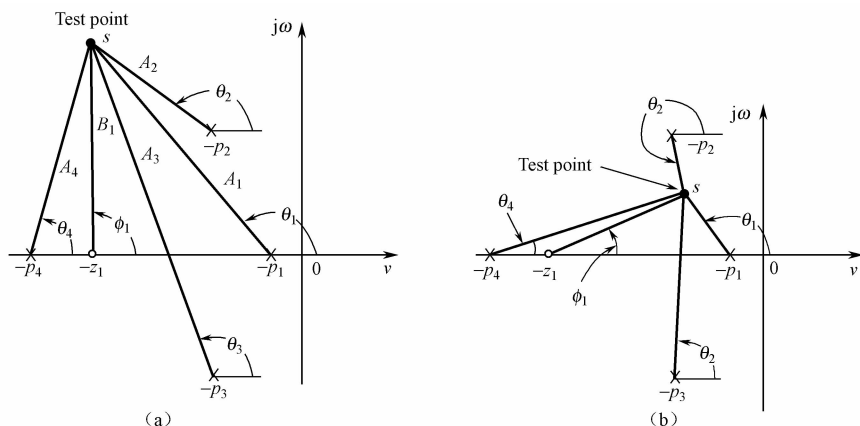


Figure 4.7.2 (a) and (b) diagrams showing angle measurements from open loop poles and open-loop zero to test point s .

Note that, because the open-loop complex-conjugate poles and complex-conjugate zeros, if any, are always located symmetrically about the real axis, the root loci are always symmetrical with respect to this axis^[6]. Therefore, we only need to construct the upper half of the root loci and draw the mirror image of the upper half in the lower-half s plane.

4.7.2 Specialized English Words

Root-Locus 根轨迹

transient response 瞬态响应, 暂态响应

loop gain 环增益, 环路增益

closed-loop pole 闭环极点

performance specifications 性能指标, 性能参数

root loci 根轨迹(复数)

damping ratio 阻尼比, 阻尼率

dominant pole 主导极点, 主极点

magnitude condition 模值条件

denominator 分母

complex quantity 复变量

complex plane 复平面

complex-conjugate 复共轭, 复共轭的

4.7.3 Notes

[1] It is important, therefore, that the designer knows how the closed-loop poles move in the s plane as the loop gain is varied. 这个复合句中, 主句“*It is important*”的实际主语为“*that*”引导的从句“*that the designer knows how the closed-loop poles move in the s plane as the loop gain is varied*”, “*It*”只是形式主语。主语从句中, “*the designer*”为主语, “*knows*”为谓语动词, “*how the closed-loop poles move in the s plane as the loop gain is varied*”则是“*knows*”的宾语从句。注意这个宾语从句中“*as the loop gain is varied*”又是“*move*”的状语从句。可见本句在结构上可分成四个层级。全句可译为“因此, 当环增益变化时, 设计者知道闭环极点在 s 平面如何移动, 是十分重要的。”

[2] The root locus is the locus of roots of the characteristic equation of the closed-loop system as a specific parameter (usually, gain K) is varied from zero to infinity, giving the method its name. 在这个复合句中,“The root locus is the locus of roots of the characteristic equation of the closed-loop system”为主句,“as a specific parameter (usually, gain K) is varied from zero to infinity”为条件状语从句。现在分词短语“giving the method its name”则是整句的状语。全句可译为“根轨迹是当某一参数(通常是增益 K)从零到无穷之间变化时,闭环系统特征方程的根的轨迹,该方法由此得名。”

[3] In designing a linear control system, we find that the root-locus method proves quite useful since it indicates the manner in which the open-loop poles and zeros should be modified so that the response meets system performance specifications. 在这个复合句中,“we”是主句,“find”是谓语动词,“that”以下为“find”的宾语从句。从句的结构较复杂:“the root-locus method proves quite useful”为从句的主要部分,“since it indicates the manner”是从句的状语从句,表示原因;而状语从句的宾语“the manner”带有一个定语从句“in which the open-loop poles and zeros should be modified”;最后,这个定语从句又有自己的状语从句“modified so that the response meets system performance specifications”。这样,本句在结构上就多达五个层级,是不多见的。全句可译为“在设计线性控制系统时,我们发现根轨迹法非常有用,它指明了如何通过修正开环极点、零点,使得响应满足系统的性能指标。”

[4] Then the root loci for the system are the loci of the closed-loop poles as the gain K is varied from zero to infinity. 本复合句结构不复杂,“Then the root loci for the system are the loci of the closed-loop poles”为主句,“as the gain K is varied from zero to infinity”为状语从句。全句可译为“那么当增益 K 从零到无穷变化时,系统的根轨迹就是闭环极点的轨迹。”

[5] Remember that the angles of the complex quantities originating from the open-loop poles and open-loop zeros to the test point are measured in the counterclockwise direction. 这是一祈使句,省略了主语“You or we (must)”,“Remember”为谓语动词,“that”以下为“Remember”的宾语从句。从句的主语为“the angles”,“of the complex quantities originating from the open-loop poles and open-loop zeros to the test point”是其定语,“are measured”为谓语。全句可译为“记住,复平面的相角是指从开环零点和开环极点到测试点 s 沿逆时针方向所测得的值。”

[6] Note that, because the open-loop complex-conjugate poles and complex-conjugate zeros, if any, are always located symmetrically about the real axis, the root loci are always symmetrical with respect to this axis. 和上面注释的一样,本句也是祈使句结构,只不过从句稍为复杂一些,即从句本身是个主从式复合结构,其中“the root loci are always symmetrical with respect to this axis”为主要部分,而“because the open-loop complex-conjugate poles and complex-conjugate zeros, if any, are always located symmetrically about the real axis”为原因从句。注意“if any”为插入的条件状语,意为“只

要有(开环共轭复极点和共轭复零点)的话”。全句可译为“请注意,由于只要存在开环共轭复极点和共轭复零点,它们总是关于实轴对称的,所以根轨迹也总是关于实轴对称的。”

4.7.4 Reference Translation

根 轨 迹

闭环系统的暂态响应的基本特征与闭环极点的位置密切相关。如果该系统的环增益是变化的,那么闭环极点的位置取决于所选环增益的值。因此,当环增益变化时,设计者知道闭环极点在 s 平面如何移动,是十分重要的。

从设计的观点来说,在一些系统下简单地进行增益调整就可以使闭环极点移动到所需要的位置,这样,设计问题就转变为对合适增益值的选择。如果仅调整增益并不能满足设计要求,就有必要增加系统补偿器。

闭环极点是特征方程的根。求解三阶以上的特征方程的根是很费力的,并且需要借助于计算机(MATLAB 为该问题提供了一个简便的方法)。然而,仅仅找出特征方程的根是不够的,因为当开环传递函数的增益发生变化时,特征方程也会随之变化,计算就必须重新进行。

W. R. Evans 提出了一种寻找闭环特征方程根的简单方法,并在控制工程中得到广泛应用。这种方法称为根轨迹法,即画出特征方程对应于所有系统参数值的根。这样,在所绘出的图上就可以确定与每个参数值对应的根的位置。注意,一般情况下该参数就是增益,但是也有可能是开环传递函数的任何其他变量。如果没有其他说明,我们假定开环传递函数的增益是一个可以从零到无穷大变化的参数。

运用根轨迹法,通过改变增益值以及增加开环零点,设计者能够推断它们对闭环极点的位置的影响。因此,不管是手工绘图还是运用计算机软件如 MATLAB 绘图,都要求设计者对描绘闭环系统的根轨迹法有较深的理解。

根轨迹法:根轨迹法的基本思想是使传递函数围绕 -1 的 s 的值必须满足系统的特征方程。

根轨迹是当某一参数(通常是增益 K)从零到无穷之间变化时,闭环系统特征方程的根的轨迹,该方法由此得名。这种图清楚地显示出每一个开环极点、零点对闭环极点的位置的影响。

在设计线性控制系统时,我们发现根轨迹法非常有用,它指明了如何通过修正开环极点、零点,使得响应满足系统的性能指标。这种方法特别适合快速获得近似结果。

由于运用 MATLAB 绘制根轨迹非常简单,有人会认为手工绘制根轨迹图是在浪费时间和精力。然而,用手工绘制根轨迹图的经验对理解计算机生成根轨迹图,以及快速地获得根轨迹的粗略含义是非常有价值的。

通过运用根轨迹法,可以确定环增益值 K ,使得闭环主导极点的阻尼率符合预期值。如果系统中的开环极点、零点的位置是系统变量,那么根轨迹法提供了选择开环极点、零点位置的方法。

相角和模值条件:考虑图 4.7.1,其开环传递函数为

$$\frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)} \quad (4.7.1)$$

令式(4.7.1)右边分母为零,可以得到闭环系统的特征方程,即

$$1 + G(s)H(s) = 0$$

或

$$G(s)H(s) = -1 \quad (4.7.2)$$

这里我们假定 $G(s)H(s)$ 是关于 s 的多项式,由于 $G(s)H(s)$ 是复变量,方程(4.7.2)可分解为以下两个方程——相角和模值方程,即

相角条件:

$$\angle G(s)H(s) = \pm 180^\circ(2k+1), k = 0, 1, 2, \dots \quad (4.7.3)$$

模值条件:

$$|G(s)H(s)| = 1 \quad (4.7.4)$$

同时满足相角和模值条件的 s 的值是特征方程的根,亦即闭环极点。在复平面中只要满足相角条件的点的轨迹就是根轨迹。对应于某个给定增益值的特征方程的根(闭环极点)可以由模值条件确定。

在许多情况下, $G(s)H(s)$ 中含有增益参数 k ,从而特征方程可以写成

$$1 + \frac{k(s+z_1)(s+z_2)\cdots(s+z_m)}{(s+p_1)(s+p_2)\cdots(s+p_n)} = 0$$

那么当增益 K 从零到无穷变化时,系统的根轨迹就是闭环极点的轨迹。

请注意,由根轨迹法绘制系统的根轨迹时,我们必须知道 $G(s)H(s)$ 的零、极点的位置。记住,复平面的相角是指从开环零点和开环极点到测试点 s 沿逆时针方向所测得的值。例如,如果 $G(s)H(s)$ 给定如下:

$$G(s)H(s) = \frac{k(s+z_1)}{(s+p_1)(s+p_2)(s+p_3)(s+p_4)}$$

式中 $-p_2$ 和 $-p_3$ 是复共轭极点,那么 $G(s)H(s)$ 的相角为

$$\angle G(s)H(s) = \phi_1 - \theta_1 - \theta_2 - \theta_3 - \theta_4$$

其中, ϕ_1 , θ_1 , θ_2 , θ_3 和 θ_4 均按逆时针旋转方向测量,如图 4.7.2(a)和(b)所示。系统的 $G(s)H(s)$ 的模值为

$$|G(s)H(s)| = \frac{kB_1}{A_1A_2A_3A_4}$$

式中 A_1, A_2, A_3, A_4 和 B_1 分别是复数 $s+p_1, s+p_2, s+p_3, s+p_4$ 和 $s+z_1$ 的模值,如图 4.7.2(a)所示。

请注意,由于只要存在开环共轭复极点和共轭复零点,它们就总是关于实轴对称,所

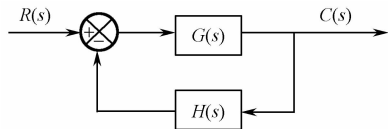


图 4.7.1 控制系统

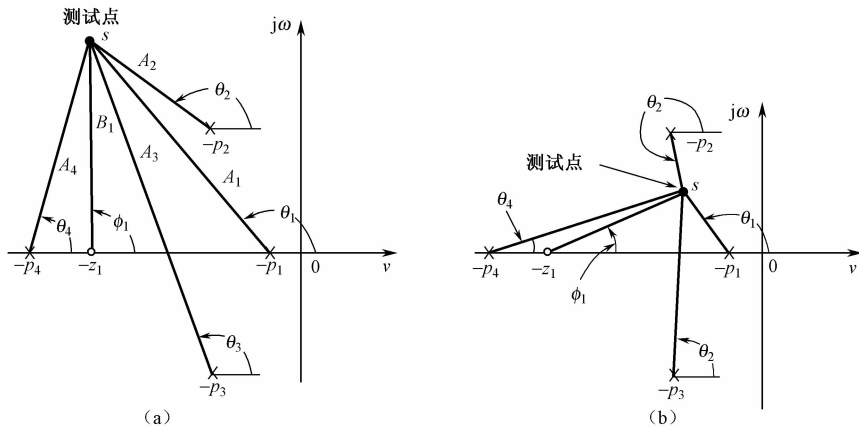


图 4.7.2 (a)和(b)显示从开环零极点到测试点 s 的相角测量

以根轨迹也总是关于实轴对称的。因此，只需要作出上半 s 平面的根轨迹，然后利用镜像对称就可以作出下半 s 平面的根轨迹。

4.7.5 Reading Materials

In control theory, the root locus is the locus of the poles and zeros of a transfer function as the system gain K is varied on some interval. The root locus is a useful tool for analyzing single input single output (SISO) linear dynamic systems. A system is stable if all of its poles are in the left-hand side of the s -plane (for continuous systems (连续系统)) or inside the unit circle of the z -plane (for discrete systems(离散系统)).

Suppose there is a motor with a transfer function expression $P(s)$, and a controller with both an adjustable gain K and a transfer function expression $C(s)$. A unity feedback loop is constructed to complete this feedback system. For this system, the overall transfer function is given by $T(s) = (KCP)/(1+KCP)$. Thus the closed-loop poles (roots of the characteristic equation) of the transfer function are the solutions to the equation $1 + KC(s)P(s) = 0$. The principal feature of this equation is that roots may be found wherever $KCP = -1$. The variability of K (that's the gain you can choose for the controller) removes amplitude from the equation, meaning the complex valued evaluation of the polynomial in s $K(s)C(s)$ needs to have net phase of 180 deg, wherever there is a closed loop pole. We are solving a root cracking problem using angles alone!

So there is no computation per-se, only geometry. The geometrical construction (几何作图) adds angle contributions from the vectors(向量) extending from each of the poles of KC to a prospective closed loop root (pole) and subtracts the angle contributions from similar vectors extending from the zeros, requiring the sum be 180.

The vector formulation arises from the fact that each polynomial term in the factored CP , $(s-a)$ for example, represents the vector from a which is one of the roots, to s which is the prospective closed loop pole we are seeking. Thus the entire polynomial is the product of these terms, and according to vector mathematics the angles add (or subtract, for terms in the denominator) and lengths multiply (or divide). So to test a point for inclusion on the root locus, all you do is add the angles to all the open loop poles and zeros. Indeed a form of protractor, the “spirule”(螺旋线尺) was once used to draw exact root loci.

From the function $T(s)$, we can also see that the zeros of the open loop system (CP) are also the zeros of the closed loop system. It is important to note that the root locus only gives the location of closed loop poles as the gain K is varied, given the open loop transfer function. The zeros of a system can not be moved.

Using a few basic rules, the root locus method can plot the overall shape of the path (locus) traversed by the roots as the value of K varies. The plot of the root locus then gives an idea of the stability and dynamics of this feedback system for different values of k .

The method is due to W. R. Evans (AIEE Transactions, 1948).

Part 5 Sensing Technology

5.1 Sensors in Manufacturing

5.1.1 Text

We now review a basic classification of sensors based upon the principle of operation. Several excellent texts exist that offer detailed descriptions of a range of sensors and these have been summarized in the material below. We distinguish here between a transducer and a sensor even though the terms are often used interchangeably. A transducer is generally defined as a device that transmits energy from one system to another, often with a change in form of the energy^[1]. A good example is a piezoelectric crystal which will output a current or charge when mechanically actuated. A sensor, on the other hand, is a device which is ‘sensitive’ to (meaning responsive to or otherwise affected by) a physical stimulus (e. g. light) and then transmits a resulting impulse for interpretation or control. Clearly there is some overlap as in the case of a piezoelectric actuator (responding to a charge and outputting a motion or force) and a piezoelectric sensor (outputting a charge for a given force or motion input). In one case, the former, the piezo device acts as a transducer and in the other, the latter, as a sensor. The terms can often be used interchangeably without problem in most cases.

A sensor, according to Webster’s Dictionary is ‘a device that responds to a physical (or chemical) stimulus (such as heat, light, sound, pressure, magnetism, or a particular motion) and transmits a resulting impulse (as for measurement or operating control)’. Sensors are in this way devices which first perceive an input signal and then convert that input signal or energy to another output signal or energy for further use. We generally classify signal outputs into six types:

- mechanical
- thermal (i. e. kinetic energy of atoms and molecules)
- electrical
- magnetic
- radiant (including electromagnetic radio waves, micro waves, etc.) and
- chemical

Sensors now exist, and are in common use, that can be classified as either ‘sensors’ on silicon as well as ‘sensors in silicon’. We shall discuss the basic

characteristics of both types of silicon ‘micro-sensors’ but introduce some of the unique features of the latter which are becoming more and more utilized in manufacturing^[2]. The small size, multi-signal capability, and signal processing and control systems make them extremely practical. In addition, as a result of their relatively low cost, these are expected to be the ‘sensors of choice’ in the future.

The six types of signal outputs listed above reflect the 10 basic forms of energy that sensors convert from one form to another. These are listed in Table 5. 1. 1. In practice, these 10 forms of energy are condensed into the six signal types listed as we can consider atomic and molecular energy as part of chemical energy, gravitational and mechanical as one, mechanical, and we can ignore nuclear and mass energy^[3]. The six signal types (hence basic sensor types for our discussion) represent ‘measurands’ extracted from manufacturing processes that give us insight into the operation of the process. These measurands represent measurable elements of the process and, further, derive from the basic information conversion technique of the sensor. That is, depending on the sensor, we will probably have differing measurands from the process. However, the range of measurands available is obviously closely linked to the type of (operating principle) of the sensor employed. Table 5. 1. 2, defines the relevant measurands from a range of sensing technologies. The ‘mapping’ of these measurand/sensing pairs on to a manufacturing process is the basis of developing a sensing strategy for a process or system. The measurands give us important information on the:

- process (the electrical stability of the process, in electrical discharge machining, for example).
- effects of outputs of the process (surface finish, dimension, for example).
- state of associated consumables (cutting fluid contamination, lubricants, tooling, for example).

Table 5.1.1 Forms of energy converted by sensors.

Energy form	Definition
Atomic	Related to the force between nuclei and electrons
Electrical	Electric fields, current, voltage, etc.
Gravitational	Related to the gravitation attraction between a mass and the Earth
Magnetic	Magnetic fields and related effects
Mass	Following relativity theory ($E = mc^2$)
Mechanical	Pertaining to motion, displacement/velocity, force, etc.
Molecular	Binding energy in molecules
Nuclear	Binding energy in electrons
Radiant	Related to electromagnetic radio waves, microwaves, infrared, visible light, ultraviolet, X-rays and C-rays
Thermal	Related to the kinetic energy of atoms and molecules

Table 5.1.2 Process measurands associated with sensor signal types.

Signal output type	Associated process measurands
Mechanical (includes acoustic)	Position (linear, angular) Velocity Acceleration Force Stress, pressure Strain Mass, density Moment, torque Flow velocity, rate of transport Shape, roughness, orientation Stiffness, compliance Viscosity Crystallinity, structural integrity Wave amplitude, phase, polarization, spectrum Wave velocity
Electrical	Charge, current Potential, potential difference Electric field (amplitude, phase, polarization, spectrum) Conductivity Permittivity
Magnetic	Magnetic field (amplitude, phase, polarization, spectrum) Magnetic flux Permeability
Chemical (includes biological)	Components (identities, concentrations, states) Biomass (identities, concentrations, states)
Radiation	Type Energy Intensity Emissivity Reflectivity Transmissivity Wave amplitude, phase, polarization, spectrum Wave velocity
Thermal	Temperature Flux Specific heat Thermal conductivity

Finally, there are a number of technical specifications of sensors that must be addressed in assessing the ability of a particular sensor/output combination to measure

robustly the state of the process^[4]. These specifications relate to the operating characteristics of the sensors and are usually the basis for selecting a particular sensor from a specific vendor, e. g. :

- ambient operating conditions
- full-scale output
- hysteresis
- linearity
- measuring range
- offset
- operating life
- output format
- overload characteristics
- repeatability
- resolution
- selectivity
- sensitivity
- response speed (time constant)
- stability/drift

5.1.2 Specialized English Words

transducer 变换器
piezoelectric crystal 压电晶体
mechanically actuated 机械压力
physical stimulus 物理刺激
piezoelectric actuator 压电传动装置,
压电驱动器
piezoelectric sensor 压电式传感器
piezo device 压电装置
magnetism 磁性,吸引力,磁学
kinetic energy 动能
radiant 辐射,辐射的,发光的
electromagnetic radio waves 电磁无
线电波
micro waves 微波
multi-signal capability 多信号处理能力
ease of integration into 易于集成到
.....

atomic 原子,原子的,原子能的
molecular energy 分子能
gravitational 引力
mass energy 质能,质能关系
electrical discharge machining 电火
花加工
surface finish 表面光洁度
cutting fluid 切削液
lubricants 润滑剂,润滑物,润滑油
tooling 模具,加工
magnetic fields 磁场
strain 应变,紧张,拉紧
relativity theory 相对论
pertaining to motion 相对运动
displacement 位移
binding energy 结合能
infrared 红外线

visible light 可见光
 ultraviolet 紫外线,紫外线的
 acceleration 加速度
 flow velocity 流速
 rate of transport 传输率
 moment 力矩,弯矩
 torque 转矩,扭矩,力矩
 roughness 表面粗糙度
 stiffness 刚度
 viscosity 黏度
 crystallinity 结晶度
 structural integrity 结构整体性
 wave velocity 波速
 amplitude 振幅,波幅

phase 相位,相
 polarization 偏振,极化
 spectrum 频谱,光谱,谱
 conductivity 电导率
 permittivity 介电常数,电容率
 magnetic flux 磁通量,磁通
 permeability 渗透率,渗透性
 intensity 光强,强度
 emissivity 发射率,辐射率
 reflectivity 反射率,反射,反射系数
 transmissivity 透射率,透过率
 specific heat 比热容
 thermal conductivity 热导率

5.1.3 Notes

[1] A transducer is generally defined as a device that transmits energy from one system to another, often with a change in form of the energy. 此句的主句是“A transducer is defined as a device”。句中“that”引导定语从句,修饰先行词“device”。全句可翻译为“一般将变换器定义为从一个系统向另一个系统传输能量的设备,往往伴随着能量形式的变化。”

[2] We shall discuss the basic characteristics of both types of silicon ‘micro-sensors’ but introduce some of the unique features of the latter which are becoming more and more utilized in manufacturing. 句中“which”引导定语从句,修饰“features”。全句可译成“我们讨论这两种类型的硅‘微传感器’的基本特征,但对在制造业中应用变得越来越多的后者,我们还要介绍一些它独有的特点。”

[3] In practice, these 10 forms of energy are condensed into the six signal types listed as we can consider atomic and molecular energy as part of chemical energy, gravitational and mechanical as one, mechanical, and we can ignore nuclear and mass energy. 句中第一个“as”为连接词,引导原因状语从句“we ... energy”;第二个“as”为介词,意为“视为”。全句可译为“在实践中,可以将原子与分子能视为化学能的一部分,可以将引力和机械能统一为机械能,还可以忽略原子能和质能,这样一来,这 10 种能量就可以压缩成所列出的六类信号”。

[4] Finally, there are a number of technical specifications of sensors that must be addressed in assessing the ability of a particular sensor/output combination to measure robustly the state of the process. 此句的主句是“there are a number of technical specifications of sensors”。句中“that”引导定语从句“must ... process, “technical”是先

行词。全句可译为“最后,在评估某种‘传感-输出组合’是否具备对制造过程的状态进行稳定的测量的能力时,有一系列技术参数必须加以考虑。”

5.1.4 Reference Translation

传感器在制造业中的应用

现在我们按工作原理对传感器的基本分类做一个评估。几篇非常优秀的文章对各种传感器做了详细的介绍,这在下面的材料中做了总结。在这里,我们要对变换器和传感器做一区别,尽管这些词通常可以互换使用。一般将变换器定义为从一个系统向另一个系统传输能量的设备,往往伴随着能量形式的变化。压电晶体就是一个很好的例子。受到机械压力作用时,压电晶体将输出电流或电荷。传感器则是一种对物理刺激(如光)十分“敏感”(指做出反应或受到影响),并能传送所产生的脉冲用于显示或控制的装置。显然存在着一些重叠的情况,如压电传动装置(对电荷做出反应并输出运动或力)和压电式传感器(根据所给的力或运动输入输出电荷)。一种情况下,即前者,压电装置就是一个变通器;另一种情况下,即后者,则是一个传感器。大多数情况下,两个词互换使用是没有问题的。

根据韦氏字典的解释,传感器是“对物理(或化学)的刺激(如热、声、光、压力、磁性或某一特定动作)做出反应并传送所产生的脉冲(如测量或操纵装置)的一种装置”。传感器装置是这样工作的:首先探测输入信号,然后将输入信号或能量转换成另一种可以进一步利用的输出信号或能量。我们一般把信号输出分为六类:

- 机械
- 热(即原子和分子的动能)
- 电气
- 磁性
- 辐射(包括电磁无线电波、微波等)
- 化学

现在,传感器不但有而且使用广泛。可以把传感器分为“硅片上的传感器”和“硅制成的传感器”两大类。我们讨论这两种类型的硅“微传感器”的基本特征。但对在制造业中应用变得越来越多的后者,我们还要介绍一些它独有的特点。体积小、多信号处理能力、并且易于集成到信号处理和控制系统之中,这些特点使得它们非常实用。此外,它们的成本也相当低,这些都使得这种传感器有望成为未来的“首选传感器”。

上述六种类型的输出信号反映了能量从一种形态转换成另一种形态时的 10 种基本形式(如表 5.1.1 所示)。在实践中,可以将原子与分子能视为化学能的一部分,可以将引力和机械能统一为机械能,还可以忽略原子能和质能,这样一来,这 10 种能量就可以压缩成所列出的六类信号。这六类信号(也是我们讨论的传感器的基本类型)代表了从制造过程中获得的‘被测量’,从中我们可以深入的了解制造过程的运作情况。这些被测量代表了制造过程中的可测量要素,更和传感器信息转换基本原理相关。这就是说,采用不同的传感器,我们大概能获得关于制造过程中的不同被测量。然而,采用哪些被测量很显然是

与所用传感器类型(工作原理)密切相关的。表 5.1.2 定义了不同的传感器信息类型及被测量,这些被测量传感器对与制造过程的“映射”是过程或系统开发传感策略的基础。被测量提供了以下重要信息:

- 制造过程(如电火花加工过程的电稳定性)。
- 制造过程所产生的影响(如表面光洁度、尺寸)。
- 相关消耗品的状态(如切削液的污染、润滑剂、模具)。

表 5.1.1 传感器转换的能量形式

能 量 形 式	定 义
原子	与原子核和电子之间的力有关
电气	电场、电流、电压等
引力	涉及物质和地球之间的吸引力
力	磁场和相关的影响
质量	依据相对论(为 $E = mc^2$)
机械	相对运动、位移/速度、外力等
分子	分子的结合能
核	电子的结合能
辐射	与电磁波、微波、红外线、可见光、紫外线、X 射线和 C 射线相关
热	与原子和分子的动能相关

表 5.1.2 被测量与传感信息类别

信号输出形式	相关被测量	信号输出形式	相关被测量
机械(包括声波)	位置(线性,角度)	磁	磁场(振幅,相位,偏振,频谱)
	速度		磁通量
	加速度		渗透率
	力	化学(包括生物)	成分(特性,浓度,状态)
	应力、压力		生物量(特性,浓度,状态)
	应变	辐射	类型
	质量、密度		能量
	力矩、转矩		光强
	流速、传输率		发射率
	形状、表面粗糙度、方向性		反射率
	刚度、可塑性		透射率
	黏度		波振幅,相位,偏振,频谱
	结晶度、结构整体性		波速
	波振幅、相位、偏振、频谱		
	波速		
电	电荷、电流	热	温度
	电位、电位差		流量
	电场(振幅,相位,偏振,频谱)		比热容
	电导率		热导率
	介电常数		

最后,在评估某种“传感-输出组合”是否具备对制造过程的状态进行稳定测量的能力时,有一系列技术参数必须加以考虑。这些参数涉及传感器的运行特性,也是决定是否从某个特定的供应商那里选用某种特定的传感器的基础。例如:

- 工作的环境条件
- 满标输出
- 滞后
- 线性
- 测量范围
- 偏移量
- 工作寿命
- 输出形式
- 过载特性
- 可重复性
- 分辨率
- 选择性
- 敏感性
- 响应速度(时间常数)
- 稳定性/漂移性

5.1.5 Reading Materials

A sensor is a device that measures a physical quantity and converts it into a signal which can be read by an observer or by an instrument. For example, a mercury thermometer(水银温度计) converts the measured temperature into expansion and contraction of a liquid which can be read on a calibrated glass tube. A thermocouple converts temperature to an output voltage which can be read by a voltmeter(电压表). For accuracy, all sensors need to be calibrated against known standards.

Sensors are used in everyday objects such as touch-sensitive elevator buttons and lamps which dim or brighten by touching the base. There are also innumerable applications for sensors of which most people are never aware. Applications include automobiles, machines, aerospace, medicine, industry, and robotics.

A sensor's sensitivity indicates how much the sensor's output changes when the measured quantity changes. For instance, if the mercury in a thermometer moves 1 cm when the temperature changes by 1°, the sensitivity is 1 cm/1°. Sensors that measure very small changes must have very high sensitivities.

Technological progress allows more and more sensors to be manufactured on a microscopic(微观) scale as micro sensors using MEMS technology. In most cases, a micro sensor reaches a significantly higher speed and sensitivity compared with macroscopic(宏观的) approaches. See also MEMS sensor generations.

A good sensor obeys the following rules:

- the sensor should be sensitive to the measured property
- the sensor should be insensitive to any other property
- the sensor should not influence the measured property

Ideal sensors are designed to be linear. The output signal of such a sensor is linearly proportional (成线性关系) to the value of the measured property. The

sensitivity is then defined as the ratio between output signal and measured property. For example, if a sensor measures temperature and has a voltage output, the sensitivity is a constant with the unit [V/K]; this sensor is linear because the ratio is constant at all points of measurement.

5.2 Micro Thermocouples

5.2.1 Text

Unlike the metal and semiconducting resistors, a thermocouple is a potentiometric

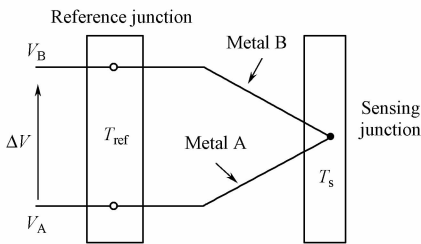


Figure 5. 2. 1 Basic configuration of a thermocouple temperature sensor (a type of potentiometric thermal sensor).

temperature sensor in that an open circuit voltage V_T appears when two different metals are joined together with the junction held at a temperature being sensed T_s and the other ends held at a reference temperature T_{ref} (see Figure 5. 2. 1)^[1].

The basic principle is known as the Seebeck effect in which the metal have a different thermoelectric power or Seebeck coefficient $P^{[2]}$; the thermocouple is conveniently a linear devices, with the voltage output (at zero current) being given by

$$V_T = (V_B - V_A) = (P_B - P_A)(T_s - T_{ref}) = (P_B - P_A)\Delta T \tag{5. 2. 1}$$

Thermocouples are also widely used to measure temperature, and their properties are defined in British and US standards for different composition of metal and alloys, for example, types B, E, J, K, N, R, S, and T. Typically, they can operate from -100 to $+2000^{\circ}\text{C}$ with accuracy of between 1 and 3 percent for a full-scale operation (FSO) ^[3].

Here, we are mainly interested in whether a temperature sensor can be interested in a silicon process to become either a temperature micro sensor or part of a silicon-based MEMS device. Table 5. 2. 1 summaries the typical properties of conventional temperature sensor and more importantly, whether they can be integrated into a standard integrated circuit (IC) process.

As is apparent from Table 5. 2. 1, it is possible to integrate resistive temperature sensors such as the platinum Pt-100. However, the deposition of platinum or the thermistor oxide is a nonstandard IC process and therefore requires, additions pre or post-IC processing steps. The inclusion of nonstandard materials during, for example, a CMOS process, which is ‘intermediate’ CMOS, is generally regarded as highly undesirable and should be avoided if possible^[4].

Table 5.2.1 Properties of common temperature sensors and their suitability for integration.Modified from Meijer and van Herwaarden (1944).

Property	Pt resistor	Thermistor	Thermocouple	Transistor
Form of output	Resistance	Resistance	Voltage	Voltage
Operating range(°C)	Large −260 to +1000	Medium −80 to +180	Very large −270 to +3500	Medium −50 to +180
Sensitivity	Medium 0.4%/K	High 5%/K	Low 0.05%/K	High ~ 2mv/K
Linearity	Very good <±0.1K	Very nonlinear	Good <±1K	Good <±0.5K
Accuracy:				
— absolute	High over wide range	High over small range	Not possible	Medium
— differential	Medium	Medium	High	Medium
Cost to make	Medium	Low	Medium	Very low
Suitability for IC Integration	Not a standard process	Not a standard process	Yes	Yes-very easily

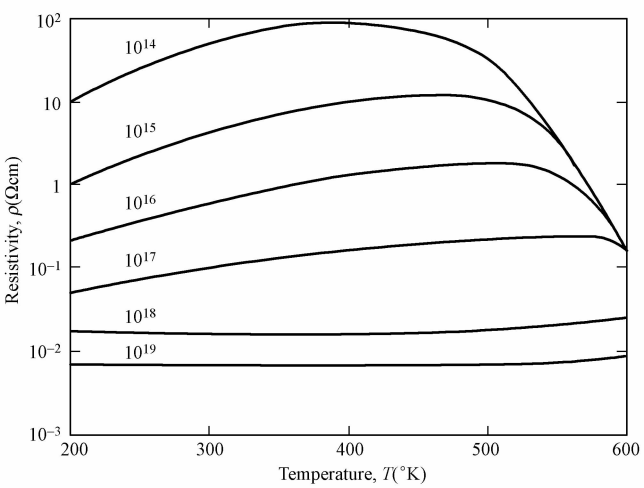


Figure 5.2.2 Temperature-dependence of single-crystal silicon doped at various levels (n-type) From Wolf (1969).

It is possible to fabricate silicon resistors in standard silicon IC process. For example, five or more resistors can be made of doped silicon in a standard bipolar process, such as a base resistor, emitter resistor, or an epi-resistor, and two or three resistors can be made in a CMOS process. The resistivity of a single crystal of silicon varies with temperature and doping level as illustrated in Figure 5.2.2, and the lightly doped silicon provided the highest TCR. In practice, it is difficult to make single-crystal silicon with an impurity level below $\sim 10^{12} \text{ cm}^{-3}$; therefore, it will not behave as an intrinsic semiconductor with a well-defined Arrhenius temperature-dependence because the intrinsic carrier concentration is about 10^{10} cm^{-3} at room temperature. In highly

doped silicon resistors ($\sim 10^{18} \text{ cm}^{-3}$), the temperature-dependence approximates reasonably well to the second-order polynomial given in Equation (5.2.1). Nevertheless, the temperature-dependence of a silicon resistor is nonlinear and depends upon

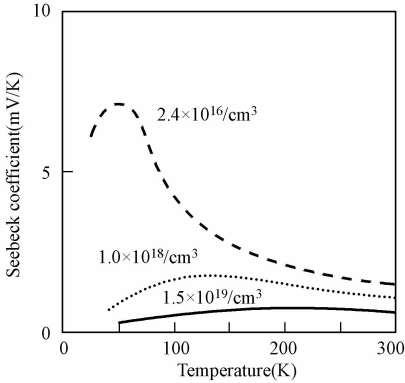


Figure 5.2.3 Variation of Seebeck coefficient for single-crystal silicon doped with temperature at different concentrations of boron (i.e. p-type). Adapted from Geballe and Hull (1955).

based thermocouple becomes more linear. As a variety of doping levels are possible in a planar IC process, a Seebeck coefficient ranging from +0.5 to +5 mV/°C is achievable.

In theory, the Seebeck coefficient of a doped semiconductor is given by

$$\begin{aligned} \text{n-type: } P_{\text{n-Si}} &= -\frac{k_B}{q} \{ [\ln(N_c/n) + 2.5] + (1 + S_n) + \phi_n \} \\ \text{p-type: } P_{\text{p-Si}} &= -\frac{k_B}{q} \{ [\ln(N_v/p) + 2.5] + (1 + S_p) + \phi_p \} \end{aligned} \quad (5.2.1)$$

where k_B is the Boltzmann's constant, q is the carrier charge, N_c and N_v are the density of states at the bottom of the conductance band and top of the valence band, n and p are the donor and acceptor concentrations, s is a parameter related to the mean free time between collision and the charge carrier energy and its value varies between -1 and $+2$ depending on whether the carrier can move freely or are trapped, and finally ϕ is a phonon drag term for the carrier. In practice, the Seebeck coefficient can be readily estimated from the silicon resistivity rather than the carrier concentrations and is simply given by

$$P_{\text{Si}} \approx \frac{mk_B}{q} \ln \frac{\rho}{\rho_0} \quad (5.2.2)$$

where m is a dimensionless constant (negative for n-type and positive for p-type) and is typically around 2.6 and ρ_0 is a resistivity constant of $5 \times 10^{-6} \Omega \text{m}$.

the exit doping level, making it less suitable for use as a temperature sensor than other types of device. Therefore, the preferred approaches are to make a micro thermocouple out of silicon or, better still, use the inherent temperature sensitivity of a silicon diode or transistor.

The Seebeck coefficient of single-crystal silicon varies with both temperature and doping concentration (p-type) as shown in Figure 5.2.3. Doping has the effect of reducing the temperature variation of the coefficient itself; hence, the response of a silicon-

Therefore, a silicon thermocouple can be made in an IC process with doped silicon and a standard metal contact, for example, aluminum. Figure 5. 2. 4 shows such a thermal micro sensor and consists of a series of N identical p-Si/Al thermocouples.

The theoretical voltage output V_{out} of this thermopile is given subsequently (from Equation (5. 2. 3)) and agrees well with experimental values.

$$V_T = N(V_{p-Si} - V_{Al}) = N(P_{p-Si} - P_{Al})\Delta T \tag{5. 2. 3}$$

As the absolute Seebeck coefficient of p-type silicon is positive (e. g. $+1\text{ mV/K}$ for a sheet resistance of $200\text{ }\Omega/\text{sp}$ at 300 K) and that for aluminum is negative (i. e. $-1.7\text{ }\mu\text{V/K}$ at 300 K), an output on the order of n millivolts per degree can be achieved from a thermopile.

Polysilicon/gold

thermocouples have also been made with an output of about $+0.4\text{ mV/K}$ in which the n-type (phosphorous) polysilicon has a lower Seebeck coefficient of $-176\text{ }\mu\text{V/K}$ (for a sheet resistance of $100\text{ }\Omega/\text{sp}$ at 300 K) and the gold has a standard value of $+194\text{ }\mu\text{V/K}$. However, these are not standard IC process materials and so polysilicon-based thermocouples are not preferred fabrication route for low-cost temperature microsensors.

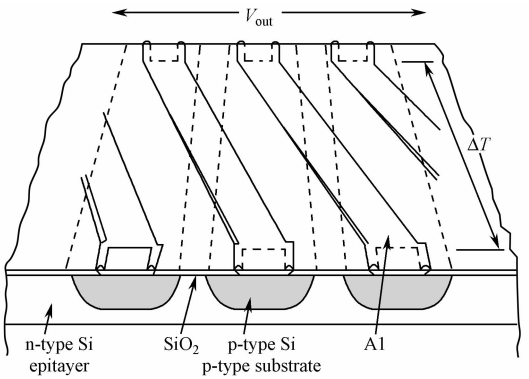


Figure 5. 2. 4 Example of a temperature micro sensor; a p-Si/Al thermopile integrated in an n-type epilayer employing a standard bipolar process. From Meijer and van Herwaarden (1944).

5. 2. 2 Specialized English Words

semiconducting resister	半导体电阻	silicon process	硅工艺加工, 硅工艺
thermocouple	热电偶	standard integrated circuit (IC)	标准集成电路 (IC)
potentiometric temperature sensor		process	标准集成电路 (IC) 工艺
电位型温度传感器		nonstandard materials	非标准材料
open circuit voltage	开路电压	Pt resistor	铂电阻
thermoelectric power	热电势, 温差	thermistor	热敏电阻
电势, 热电功率		transistor	晶体管
linear devices	线性器件, 线性装置	operating range	测量范围
full-scale operation	满量程	sensitivity	灵敏度

doped silicon 掺杂硅	planar IC process 平面集成电路工艺
standard bipolar process 标准双极工艺	conductance band 导带
base resistor 基极电阻	valence band 价带
emitter resistor 发射极电阻	phonon drag term 声子曳引项
epi-resistor 外延电阻	carrier charge 载流子电荷
TCR 电阻温度系数	resistivity constant 电阻率常数
intrinsic semiconductor 本征半导体	boron 硼
inherent temperature sensitivity 固有的温度敏感性	theoretical voltage 理论电压
silicon diode 硅二极管	polysilicon 多晶硅
doping concentration 掺杂浓度	n-type epilayer n型外延电阻

5.2.3 Notes

[1] Unlike the metal and semiconducting resistors, a thermocouple is a potentiometric temperature sensor in that an open circuit voltage V_T appears when two different metals are joined together with the junction held at a temperature being sensed T_s and the other ends held at a reference temperature T_{ref} . 此句是一个成分比较复杂的主从句结构。主句是“a thermocouple is a potentiometric temperature sensor”，介词短语“Unlike”引导状语“the metal and semiconducting resistors”，“in that”（相当于“in which”）引导原因状语定语从句，修饰“temperature sensor”。“when”引导时间状语从句“two different metals are joined together with the junction held at a temperature being sensed T_s and the other ends held at a reference temperature T_r ”，其中“held at a temperature being sensed T_s ”为过去分词短语做定语，修饰“the junction”，而“being sensed T_s ”则是现在分词短语（被动式）做定语，修饰“temperature”。“and the other ends held at a reference temperature T_{ref} ”中的“held at a reference temperature T_{ref} ”也是定语，修饰“the other ends”。全句可译为“与金属和半导体电阻不同，热电偶是一种电位型的温度传感器。当两种不同的金属连接在一起，节点一端温度为 T_s ，另一端为参考温度 T_{ref} 时，就会有开路电压 V_T 输出（见图 5.2.1）。”

[2] The basic principle is known as the Seebeck effect in which the metal have a different thermoelectric power or Seebeck coefficient P ；句子的主句是“The basic principle is known as the Seebeck effect”，“in which”引导定语从句“the metal have a different thermoelectric power or Seebeck coefficient P ”，修饰“the Seebeck effect”。全句可译为“金属具有不同的热电势或塞贝克系数 P ，称为塞贝克效应，它是热电偶的基本工作原理”。

[3] Typically, they can operate from -100°C to $+2000^{\circ}\text{C}$ with accuracy of between 1 and 3 percent for a full-scale operation (FSO). 句中介词短语“with accuracy of between 1 and 3 percent for a full-scale operation (FSO)”做宾语补足语状语。全句可译

为“通常情况下,热电偶的测量范围为 -100°C 至 $+2000^{\circ}\text{C}$,精度为满量程的 $1\%\sim 3\%$ ”。

[4]The inclusion of nonstandard materials during, for example, a CMOS process, which is ‘intermediate’ CMOS, is generally regarded as highly undesirable and should be avoided if possible. 句中“which is ‘intermediate’ CMOS”是关系代词,系“which”引导的定语从句。“during, for example, a CMOS process”=“for example, during a CMOS process”。全句可译为“在加工工艺中加入非标准材料,如在 CMOS 工艺中加入非标准材料,这称为‘中间’CMOS,被普遍认为是极不可取并应尽量避免的。”

5.2.4 Reference Translation

微型热电偶

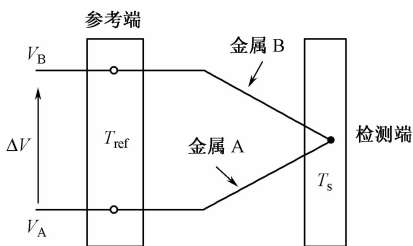


图 5.2.1 热电偶温度传感器的基本配置(电位型热传感器)

与金属和半导体电阻不同,热电偶是一种电位型的温度传感器。当两种不同的金属连接在一起,节点一端温度为 T_s ,另一端为参考温度 T_{ref} 时,就会有开路电压 V_T 输出(见图 5.2.1)。

金属具有不同的热电势或塞贝克系数 P ,称为塞贝克效应,它是热电偶的基本工作原理;热电偶是一种使用方便的线性器件,其输出电压(零电流时)由下式给出:

$$V_T = (V_B - V_A) = (P_B - P_A)(T_s - T_{ref}) = (P_B - P_A)\Delta T \tag{5.2.1}$$

热电偶也被广泛用于测量温度,以英美标准定义了多种类型的具有不同性能的热电偶,例如 B 型、E 型、J 型、K 型、N 型、R 型、S 型和 T 型,它们由不同成分的金属和合金构成。通常情况下,热电偶的测量范围为 $-100^{\circ}\text{C}\sim 2000^{\circ}\text{C}$,精度为满量程的 $1\%\sim 3\%$ 。

这里,我们主要感兴趣的是温度传感器是否可以经过硅工艺加工成为一种温度微传感器或以硅为基础的微机电系统(MEMS)中的一个器件。表 5.2.1 概括了常规下温度传感器的典型性质,更重要的是,它们是否可以由标准的集成电路(IC)工艺进行集成。

如表 5.2.1 所示,整合成电阻型温度传感器是可能的,如铂金 Pt-100。然而,白金的沉积或热氧化是一种非标准的集成电路工艺,因此,需要在集成电路加工前或加工后增补一些步骤。在加工工艺中,如在 CMOS 工艺中,加入非标准材料,称为“中间”CMOS,这被普遍认为是极不可取并应尽量避免的。

表 5.2.1 常见温度传感器的性能及集成的适应性。改编自 meijer 及 Van herwaarden(1944)

性 能	铂 电 阻	热 敏 电 阻	热 电 偶	晶 体 管
输出形式	电阻	电阻	电压	电压
测量范围(℃)	大, -260 到 +1000	中等, -80 到 +180	非常大, -270 到 +3500	中等, -50 到 +180
灵敏度	中等 0.4%/K	高 5%/K	低 0.05%/K	高 ~ 2mV/K
线性度	非常好 <±0.1K	严重非线性	好 <±1K	好 <±0.5K
精度:				
绝对精度	高(大范围内)	高(小范围内)	无	中等
差分精度	中等	中等	高	中等
成本	中等	低	中等	极低
是否适合集成	非标准工艺	非标准工艺	是	是, 极易

可以采用标准硅集成电路工艺制造硅电阻。比如,采用标准双极工艺可以把掺杂硅制成 5 个或更多电阻,如作为基极电阻、发射极电阻或外延电阻。CMOS 工艺也可以制成两到三个电阻。单晶硅的电阻率随温度及掺杂度而变化,如图 5.2.2 所示,轻掺杂硅可以达到很高的电阻温度系数。在实践中,掺杂度低于 $\sim 10^{12} \text{ cm}^{-3}$ 的单晶硅很难制成;因此,它不会表现出像本征半导体一样明确的 Arrhenius 温度依赖性,因为室温下内在的载流子浓度大约是 10^{10} cm^{-3} 。高掺杂的硅电阻($\sim 10^{18} \text{ cm}^{-3}$)的温度依赖性与式(5.2.1)的二阶多项式给出的结论很接近。不过,硅电阻的温度依赖性是非线性的,取决于出口掺杂度,这使得它跟其他类型的装置比起来,更不适合用于温度传感器。因此,较好的办法是用硅制造微热电偶,或者更好的办法是利用硅二极管或晶体管固有的温度敏感性。

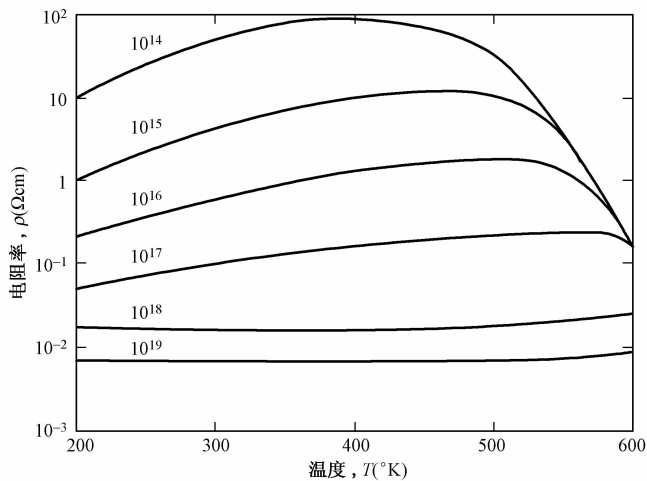


图 5.2.2 不同掺杂度单晶硅的随温度依赖性(n 型),摘自 Wolf (1969)

单晶硅的塞贝克系数会随着温度和掺杂浓度(p 型)变化,如图 5.2.3 所示。掺杂可以降低塞贝克系数本身的温度变化,因此,硅热电偶的响应的线性度得以提高。因为平面集成电路工艺可以加工各种掺杂水平的单晶硅,塞贝克系数可以在+0.5 至+5 mV/℃ 范围内变化。

理论上,掺杂半导体的塞贝克系数可由下式得出:

$$\begin{aligned} \text{n 型: } P_{\text{n-Si}} &= -\frac{k_B}{q} \{ [\ln(N_c/n) + 2.5] + (1 + S_n) + \phi_n \} \\ \text{p 型: } P_{\text{p-Si}} &= -\frac{k_B}{q} \{ [\ln(N_v/p) + 2.5] + (1 + S_p) + \phi_p \} \end{aligned} \quad (5.2.1)$$

式中, k_B 为玻尔兹曼常数, q 是载流子电荷, N_c 和 N_v 是导带底部和价带顶部自由电子密度状态, n 和 p 是施主原子和受主原子的浓度, S 是与碰撞和电荷载体的能量之间的平均自由时间有关的一个参数, 其值在 -1 到 $+2$ 之间变化, 取决于载流子是自由移动的还是被束缚的。最后, ϕ 是载流子的声子曳引项。在实践中, 塞贝克系数可以现成地从硅的电阻率而不是由载流子的浓度估计出来, 即可以简单地由下式得出:

$$P_{\text{Si}} \approx \frac{mk_B}{q} \ln \frac{\rho}{\rho_0} \quad (5.2.2)$$

式中, m 是一个无量纲的常数 (n 型为负, p 型为正), 其典型值约为 2.6 , ρ_0 是一个值为 $5 \times 10^{-6} \Omega\text{m}$ 的电阻率常数。

因此, 硅热电偶可以由集成电路工艺制成的掺杂硅与标准金属连接而成, 如铝。图 5.2.4 是这种微型传感器的示例, 它由一系列完全相同的 n 型 $p\text{-Si}/\text{Al}$ 热电偶组成。

式(5.2.3)接着给出了热电偶的理论电压输出值 V_{out} , 并与实际值相吻合。

$$V_T = N(V_{\text{p-Si}} - V_{\text{Al}}) = N(P_{\text{p-Si}} - P_{\text{Al}})\Delta T \quad (5.2.3)$$

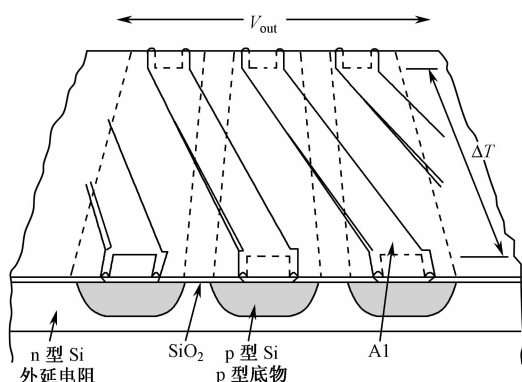


图 5.2.4 温度型微传感器示例。采用标准双极工艺集成在一个 n 型外延电阻中的 $p\text{-Si}/\text{Al}$ 热电偶。摘自 Meijer 和 van Herwaarden(1944)

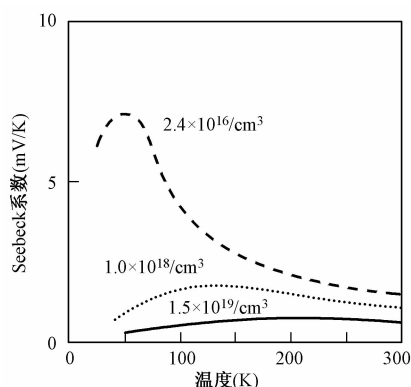


图 5.2.3 不同温度下掺杂不同浓度硼 (即 p 型) 的单晶硅的塞贝克系数的变化。
改编自 Geballe and Hull (1955)

p 型硅的绝对塞贝克系数是一个正数 (如 $200 \Omega/\text{sp}$ 的薄膜电阻温度为 300 K 时为 $+1 \text{ mV/K}$), p 型铝的绝对塞贝克系数则是一个负数 (如温度为 300 K 时为 $-1.7 \mu\text{V/K}$), 每变化 1 度, 热电偶可以输出 n 毫伏数量级的电压。当 n 型 (磷) 多晶硅有较低的塞贝克系数 $-176 \mu\text{V/K}$ (300 K 时, $100 \Omega/\text{sp}$ 的薄层电阻), 金的塞贝克系数标准值为 $194 \mu\text{V/K}$ 时, 多晶硅/金热电偶也可以获得约 $+0.4 \text{ mV/K}$ 的输出。不过, 这些都不是标准的集成电路工艺材料, 所以基于多晶硅的热电偶不适合制造低

成本的温度微传感器。

5.2.5 Reading Materials

In electronics and in electrical engineering, thermocouples are a widely used type of temperature sensor and can also be used as a means to convert thermal potential difference into electric potential difference. They are cheap and interchangeable, have standard connectors, and can measure a wide range of temperatures. The main limitation is accuracy; Kieran Thomas’ research shows that system errors of less than one degree Celsius (°C) can be difficult to achieve.

A variety of thermocouples are available, suitable for different measuring applications. They are usually selected based on the temperature range and sensitivity needed. Thermocouples with low sensitivities (B, R, and S types) have correspondingly lower resolutions. Other selection criteria include the inertness (惰性) of the thermocouple material, and whether or not it is magnetic.

Thermocouples are most suitable for measuring over a large temperature range, up to 1800 °C. They are less suitable for applications where smaller temperature differences need to be measured with high accuracy, for example the range 0 °C ~100 °C with 0.1 °C accuracy. For such applications, thermistors(热敏电阻) and resistance temperature detectors are more suitable.

5.3 Pressure Microsensors

5.3.1 Text

Pressure microsensors were the first type of silicon micromachined sensors to be developed in the late 1950s and early 1960s. Consequently, the pressure microsensors represent probably the most mature silicon micromechanical device with widespread commercial availability today. The largest market is undoubtedly the automotive, and Table 5.3.1 shows the enormous growth in the world market for automotive silicon micromachined sensors from 1989 to 1999. The two important silicon sensors are the pressure and microaccelerometer sensors, with substantial growth expected for gyrometers, which will be used for navigation^[1].

Table 5.3.1 Worldwide growth for automotive silicon micromachined sensors. From Sullivan(1993).

Year	Revenue(MEuro)	Growth-rate(%)	Year	Revenue(MEuro)	Growth-rate(%)
1989	175	—	1995	376	21
1990	283	62	1996	463	23
1991	323	14	1997	564	22
1992	321	—1	1998	679	20
1993	285	—11	1999	804.2	18
1994	312	10			

The two most common methods to fabricate pressure microsensors are bulk and surface micromachining of polysilicon. Silicon diaphragms can be made using either technique as described earlier. Figure 5. 3. 1 illustrates the basic principles of a piezo-resistive sensor and a capacitive pressure sensor.

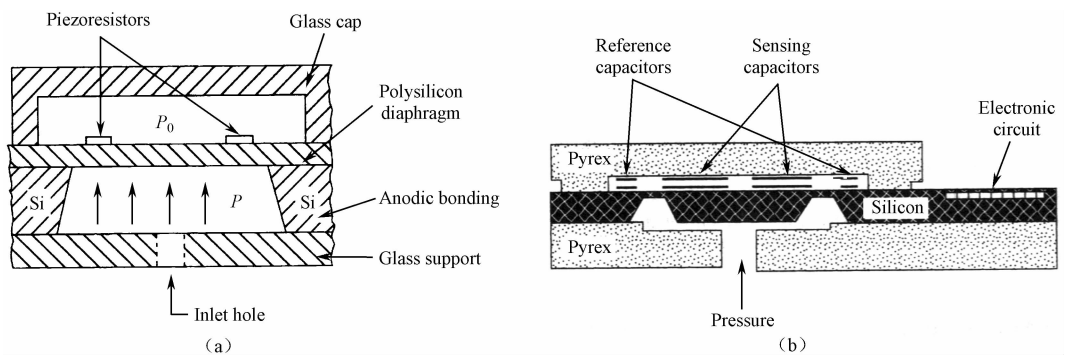


Figure 5. 3. 1 Basic types of silicon pressure sensors based on a vertical deflection; (a) piezoresistive (polysilicon) and (b) capacitive (single-crystal silicon).

The deflection in the diaphragm can be measured using piezoresistive strain gauges located in the appropriate region of maximum strain, as shown in Figure 5. 3. 1(a). The Strain gauges are usually made from doped silicon and are designed in pairs with a readout circuit such as a wheatstone bridge. The change in strain can be related to the applied pressure($P-P_0$) and stored in a lookup table. The precise relationship depends on the relevant piezoresistive coefficient Π of the diaphragm material.

$$V_{out} \propto \Delta R \propto \Pi (P - P_0) \tag{5.3.1}$$

A single crystal of silicon is a desirable material to use for the diaphragm because neither creep nor hysteresis occurs. The piezoresistive constant (Π_{44}) is typically $+138.1 \text{ pC/N}$ and that makes measuring pressure in the range of 0 to 1 MPa relatively straightforward^[2].

Figure 5. 3. 1(b) shows the general arrangement of a single-crystal pressure sensor with capacitive pickup. In this case, a capacitive bridge can be formed with two reference capacitors and the output voltage is related to the deflection of the membrane $\Delta\chi$ and hence the applied pressure ($P - P_0$)^[3].

$$V_{out} \propto \Delta C \propto \Delta\chi \propto (P - P_0) \tag{5.3.2}$$

In this case, the accurate positioning of the pickup electrodes is crucial. By controlling the background pressure P_0 , it is possible to fabricate the following basic types of pressure sensors:

- An absolute pressure sensor that is referenced to a vacuum ($P_0 = 0$)
- A gauge-type pressure sensor that is referenced to atmospheric pressure ($P_0 = 1 \text{ atm}$)

- A differential or relative type (P_0 is constant)

There are advantages and disadvantages of capacitive against piezoresistive pressure sensors and these are summarized in Table 5. 3. 2.

The main advantage of using bulk micromachining is that die electronic circuit can be more readily integrated. There are many examples of capacitive pressure sensors with digital readout. An example of a capacitive pressure sensor is shown in Figure 5. 3. 3 with a 100 μm polysilicon diaphragm and integrated capacitance circuit (Kung and Lee 1992). The output voltage from the integrated n-type metal oxide semiconductor (a MOS) circuit is also shown against air pressure in non-SI units of PSI. This design achieves a high resolution by using integrated electronics.

Table 5.3.2 Relative merits of capacitive and piezoresistive static deflection pressure sensors.

	Advantages	Disadvantage
Capacitive	More sensitive(polysilicon)	Large piece of silicon for bulk micromachining
	Less temperature-sensitive	Electronically more complicated
	More robust	Needs integrated electronics
Piezoresistive	Smaller structure than bulk capacitance	Strong temperature-dependence
	Simple transducer circuit	Piezo coefficient depends on the doping level
	No need for integration	

An alternative approach to enhance the sensitivity of silicon pressure sensors was

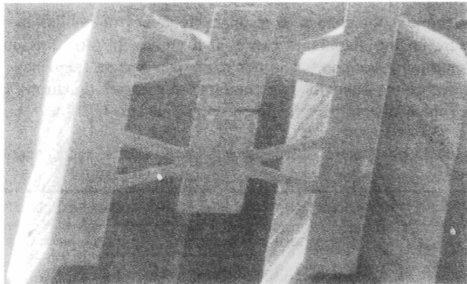


Figure 5. 3. 2 A vertical resonant capacitive pressure sensor
 based on the torsional oscillation of a strained bulk-micromachined structure. From Greenwood (1988).

Proposed by Greenwood in 1988 and comprised the use of a resonant microstructure.

Figure 5. 3. 2 shows the micromechanical structure bulk-microma-chined out of singlecry-stal silicon (Greenwood 1988).

The basic principle is the change of resonant frequency of oscillation of this structure when the pressure on the diaphragm

causes it to curve. In turn, this curvature creates tension in the shuttle mass supports and this shifts its resonant frequency. The dynamical equation that governs the behavior is a modified version of Equation (5. 3. 3) to include a tension term, which affects the effective spring constant $K_m^{[4]}$. The resonant (torsional) pressure sensor proved to have excellent resolution (a few centimeters in air) and stability (parts per million (ppm) per year) through the running of the resonator in a partial vacuum. Accordingly, it is possible to achieve a high mechanical Q factor, here about 18000 at a pressure of approximately 1 Pa, and hence achieve very high pressure sensitivities.

$$m\ddot{x} + b_m\dot{x} + k_mx = F_x(t) \quad (5.3.3)$$

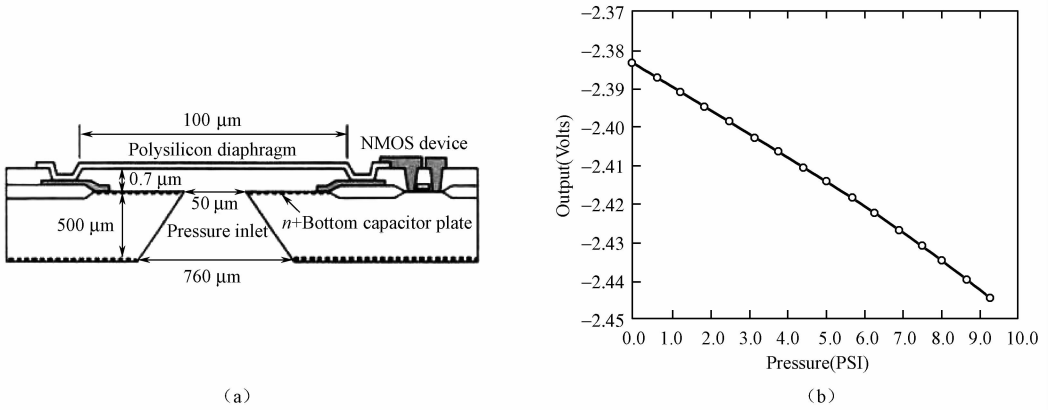


Figure 5.3.3 Polysilicon capacitive pressure sensor: (a) cross section with integrated electronics, (b) voltage response from a 100 μm square diaphragm of thickness 1 μm. From Kung and Lee(1992).

Further efforts have been made to fabricate a lateral resonant capacitive sensor employing thin film polysilicon technology. Figure 5.3.4(a) shows a resonant capacitive sensor fabricated in polysilicon along with its response (Figure 5.3.4(b)). The nonlinear response is fitted using a high-order polynomial and temperature effects are compensated for.

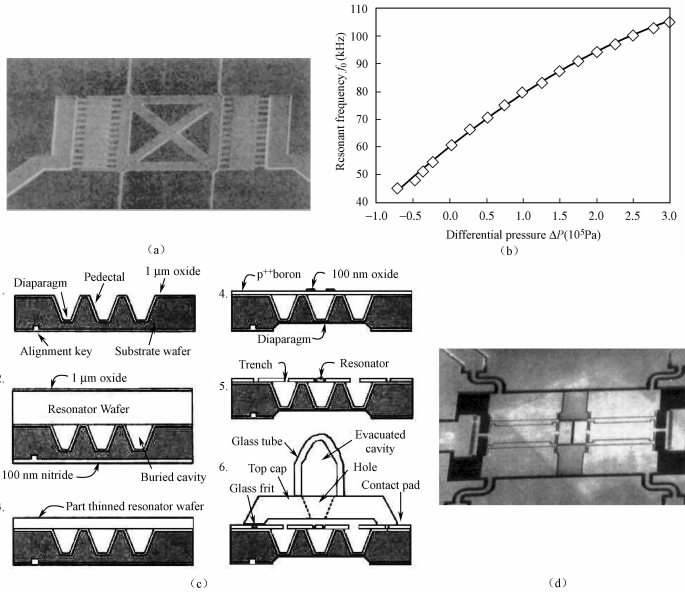


Figure 5.3.4 (a) Lateral resonant capacitive pressure sensors based on the linear oscillation of a strained surface micromachined structure; (b) its response to barometric pressure (from Welham and coworkers (1996)); (c) current silicon process; (d) latest device with piezoresistive pickup (Welham et al. 2000.)

Hence, the microstructure behaves as a nonlinear resonator and Equation (5.3.3) is extended to describe a hard spring (Duffin’s equation) so that

$$m\ddot{x} + b_m\dot{x} + k_m^0x + k_m^1x^3 = F_x(t) \tag{5.3.4}$$

Table 5.3.3 Application of silicon pressure sensors in 1997. Adapted from Madou(1997).

Application	Cost of device(Euro)	Market (MEuro)	Pressure range (kPa)	Year introduced
Manifold pressure	9	30	0—105	Current
Barometric pressure	9	100	50—105	Current
Exhaust gas recirculation	9	3.3	0—105	1989
Fuel pressure	9	97	0—105	1994
Tire pressure	n/a	455	500	1994—1995
Active suspension hydraulics	7	14	20000	1994—1995
Climate control	9	19	50—105	Current

The solution to Equation (5.3.4) is interesting because it has two possible deflections at certain frequencies. However, running the oscillator at low deflections using closed-Loop feedback avoids this stability issue. The problem with structure is that the Capacitances for drive and sensing are too low because the microshuttle is only 1 to 5 μmThick. However, recent developments of LIGA and deep RIE now make resonant lateral Structures a practical device. The resonator has now been redesigned by Welham et al. (2000) to overcome these problems together with a piezoresistive pickup. Figure 5.3.4(c) Shows the new silicon process and the fabricated device is shown in Figure 5.3.4(d). These devices have an accuracy of 0.01 percent root-mean-square (rms) or better, which, so far, exceeds that for static pressure sensors. The product is being commercialized by Druck Ltd (UK) as a precision pressure sensor because it is relatively expensive to make.

Nevertheless, the preferred technology today is bulk silicon-micromachined piezoresistive pressure sensors because of low cost, robustness and ease of circuit integration.

Table 5.3.3 summarizes the current automotive pressure sensor applications (Madou 1997).

Clearly, the automotive market for pressure sensors is enormous and commercial devices are available today from Motorola, Nova&Sensor, SSI Technologies, and other manufacturers. As costs are driven down, the move toward piezoresistive polysilicon is desirable but creates some stability and precision issues^[5]. Therefore, we may see the appearance of alternative technologies to make diaphragms such as silicon on insulator (SOI).

5.3.2 Specialized English Words

bulk 块,体,堆

micromachined 微加工的

microaccelerometer sensors 微加速度传感器
 gyrometer 陀螺测试仪,陀螺仪表
 piezoresistive sensor 压阻传感器
 deflection 应变
 strain gauges 应变片
 diaphragm 振膜,隔膜,膜片
 membrane 膜片,振片
 piezoresistive coefficient 压阻常数
 piezo coefficient 压电系数
 pickup electrodes 信号电极
 static deflection 静荷载挠度,静载挠度,静变位
 torsional oscillation 扭转振荡
 resonant frequency 共振频率
 shuttle mass support 振梁
 nonlinear resonator 非线性谐振器
 linear oscillation 线性振荡
 piezoresistive pickup 压阻传感器
 spring constant 弹簧系数
 hard spring 硬弹簧,硬质弹簧
 cross section 横截面,横切面,横剖面
 lateral 横向
 barometric pressure 气压,大气压,大气压力
 microshuttle 微型压力传感器

silicon on insulator (SOI) 绝缘硅(一种新型硅集成电路材料的简称)
 exhaust gas recirculation 排气再循环
 fuel pressure 燃料压力
 active suspension hydraulics 液压主动悬架
 piezoresistors 压敏电阻器
 anodic bonding 阳极键合
 glass support 玻璃支
 inlet hole 进入孔
 pyrex 硼硅酸玻璃
 reference capacitors 标准电容器
 oxide 氧化物
 boron 硼
 alignment key 对准标记
 substrate wafer 次级硅片
 trench 沟
 resonator 谐振器
 resonator wafer 谐振硅片
 buried cavity 隐藏孔
 hole 孔
 evacuated cavity 真空共振腔
 glass frit 玻璃粉
 contact pad 触板
 nitride 氮化物

5.3.3 Notes

[1] The two important silicon sensors are the pressure and microaccelerometer sensors, with substantial growth expected for gyrometers, which will be used for navigation. 句中“which”引导非限制性定语从句“will be used for navigation”,修饰“gyrometers”,“with substantial growth expected for gyrometers”在句中做伴随状语,其中“expected for gyrometers”为过去分词短语,修饰“growth”。全句可译为“压力传感器和微加速度传感器是两种重要的硅传感器,随着陀螺仪需求的大量增长,这两种传感器将会用于导航领域”。

[2] The piezoresistive constant (Π_{44}) is typically $+138.1 \text{ pC/N}$ and that makes measuring pressure in the range of 0 to 1 MPa relatively straight forward. 句中“and”引

导两个并列的句子。后句中的“that”指代“The piezoresistive constant (Π_{44}) is typically +138.1 pC/N”。此句可译为“压阻常数(Π_{44})的典型值为+138.1 pC/N,这就保证了压力测量值在的 0 到 1 MPa 范围内有相当好的线性度。”

[3]In this case, a capacitive bridge can be formed with two reference capacitors and the output voltage is related to the deflection of the membrane $\Delta\chi$ and hence the applied pressure ($P-P_0$). 句中中介词短语“In this case”做条件状语,“and”连接两个并列句子。此句可译为“在这种配置下,电容桥由两个参比电容组成,输出电压与振膜挠度 $\Delta\chi$ 有关,因而与所施压力($P-P_0$) 有关。”

[4] The dynamical equation that governs the behavior is a modified version of Equation (5.3.3) to include a tension term, which affects the effective spring constant K_m . 句中“which affects the effective spring constant K_m ”是“which”引导的定语从句,指代前面的句子,修饰“a tension term”。“that”引导主定语从句修饰“the dynamical equation”。此句可译为“约束振荡的动态方程是方程(5.3.3)的修正方程,增加了一个影响弹簧常数 K_m 的张力项。”

[5]As costs are driven down, the move toward piezoresistive polysilicon is desirable but creates some stability and precision issues. 句中“As costs are driven down”为原因状语从句,主句有两个并列的动词“is”和“creates”。此句可译为“由于成本的降低,压阻式多晶硅变得更加符合需求,但随之也带来了一些稳定性和精确性方面的问题。”

5.3.4 Reference Translation

微型压力传感器

微型压力传感器是 20 世纪 50 年代后期到 60 年代早期开发出来的第一种微型硅传感器。现在,微型压力传感器代表了目前最成熟、商业供应最广泛的微型硅传感器装置。它最大的市场无疑是汽车行业。表 5.3.1 显示了 1989 年到 1999 年间微型硅传感器用于汽车领域所获得的世界性突飞猛进的增长。压力和微加速度传感器是两种重要的硅传感器,随着陀螺仪需求的大量增长,这两种传感器将会用于导航领域。

表 5.3.1 世界车用微型硅传感器的增长情况

年 份	收 益 (中 欧)	增 长 率 (%)	年 份	收 益 (中 欧)	增 长 率 (%)
1989	175	—	1995	376	21
1990	283	62	1996	463	23
1991	323	14	1997	564	22
1992	321	—1	1998	679	20
1993	285	—11	1999	804.2	18
1994	312	10			

制造微型压力传感器最常见的两种微加工方法是对多晶硅进行体加工和表面加工。硅振膜也可以用前面叙述的工艺加工。图 5.3.1 说明了压阻传感器和电容式压力传感器的基本原理。

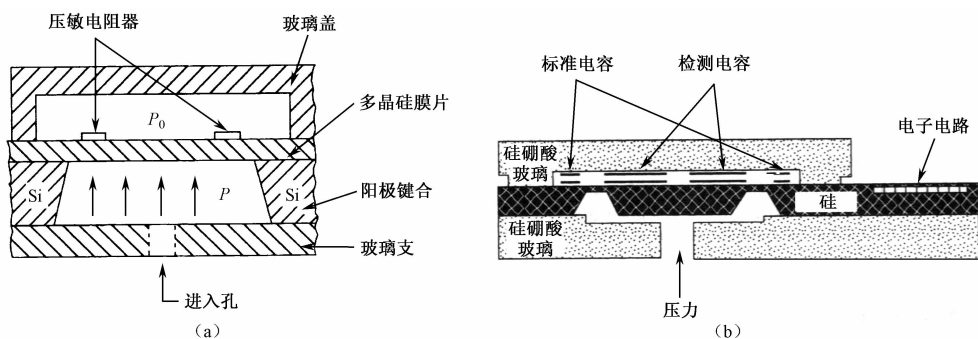


图 5.3.1 基于垂直挠度的硅压力传感器的基本类型:(a)压阻式(多晶硅)和(b)电容式(单晶硅)

如图 5.3.1 (a)所示,在最大应变的允许范围内,膜片的挠度可以用压阻式应变片来测量。应变片通常由掺杂硅制成,成对放置,带读出电路,如惠斯顿电桥。应变能根据所受压力 $(P-P_0)$ 变化并将结果存储到一个查找表中。应变与所受压力之间的精确关系取决于该膜片材料的相关压阻常数 Π 。

$$V_{\text{out}} \propto \Delta R \propto \Pi (P - P_0) \quad (5.3.1)$$

单晶硅既无蠕变,也无滞后,是制造振膜的理想材料。压阻常数(Π_{44})的典型值为 $+138.1 \text{ pC/N}$,这就保证了压力测量值在的 0 到 1 MPa 范围内有相当好的线性度。

图 5.3.1 (b)是电容电极的单晶硅压力传感器的总配置图。在这种配置下,电容桥由两个参比电容组成,输出电压与振膜挠度 $\Delta\chi$ 有关,因而与所施压力 $(P - P_0)$ 有关。

$$V_{\text{out}} \propto \Delta C \propto \Delta\chi \propto (P - P_0) \quad (5.3.2)$$

这种情况下,信号电极的准确定位是至关重要的。

通过控制环境压力 P_0 ,可以制造出以下几种基本类型的压力传感器:

- 真空绝对压力传感器($P_0 = 0$)
- 大气压型压力传感器($P_0 = 1 \text{ atm}$)
- 差分或相对型传感器(P_0 为常量)

电容式与压阻式压力传感器的优缺点对照见表 5.3.2。

表 5.3.2 电容式与压阻式静载挠度压力传感器的优缺点对照

	优 点	缺 点
电容式	更灵敏(多晶硅)	体加工需要大块硅
	温度敏感性弱	电气性能更复杂
	更具稳定性	需要集成电子技术
压阻式	比体电容的结构小	受温度影响大
	简单的变送电路	掺杂水平决定压阻系数
	无需集成	

块体电子电路更容易集成,是微加工时采用体加工方法的主要优点。关于数字量输出电容压力传感器的例子有很多。例如,图 5.3.3 所示就是一个由 $100 \mu\text{m}$ 多晶硅膜片

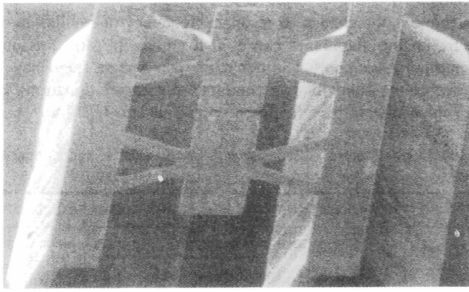
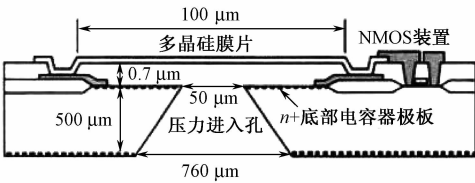


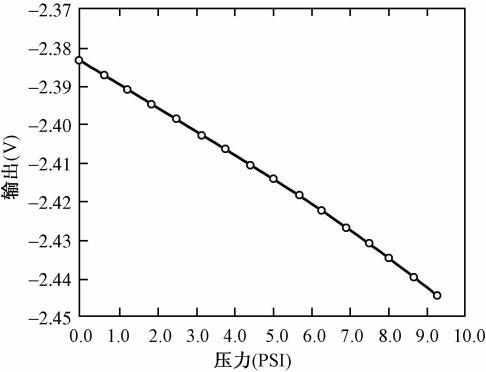
图 5.3.2 以体加工工艺制造的应变元件所产生的扭转振荡为基础的垂直谐振电容压力传感器

压力传感器检测的基本原理是：当振膜受外部压力而弯曲时，这个结构的共振频率就会发生变化。然后，弯曲引起振梁受力而紧张，共振频率就发生改变。约束振荡的动态方程是方程(5.3.3)的修正方程，增加了一个影响弹簧常数 K_m 的张力项。在局部真空环境下运行时，谐振(扭力)压力传感器有良好的分辨率(大气压下约数厘米)和稳定性(年漂移约百万分之几 ppm)。因此，实现较高的机械品质因数 Q 是可能的，压力为 1 Pa 时约 1.8 万，因此，达到了非常高的压力灵敏度。

$$m\ddot{x} + b_m\dot{x} + k_mx = F_x(t) \quad (5.3.3)$$



(a)



(b)

图 5.3.3 多晶硅电容式压力传感器：(a)集成电路的横截面；(b)厚 1 微米，面积 100 平方微米范围的振膜的电压响应。摘自 Kung 和 Lee(1992)

方程(5.3.4)的解很有趣，因为在某些频率它有两个可能的挠度。然而，振荡器在低挠度时的闭环反馈运行，可以避免不稳定问题。这种结构存在的问题是，由于这种微型传感器只有 1 至 5 微米厚，使得驱动和检测信号的电容太小。然而，随着最近 LIGA 技术和深层反应离子刻蚀(RIE)技术的发展，使得横向共振电容压力传感器成为一种实用的装置。为了解决压阻传感器的这些问题，2000 年 Welham 等人重新进行了设计。

图 5.3.4(c)展示的是其新型硅工艺,图 5.3.4(d)是制成的器件,它可以达到 0.01%的均方根(有效值)或更高的精度,这个精度目前已经超过了静态压力传感器的精度。作为一种精密的压力传感器,它的生产成本还相对较高,英国 Druck 有限公司正在努力把它的商业化。

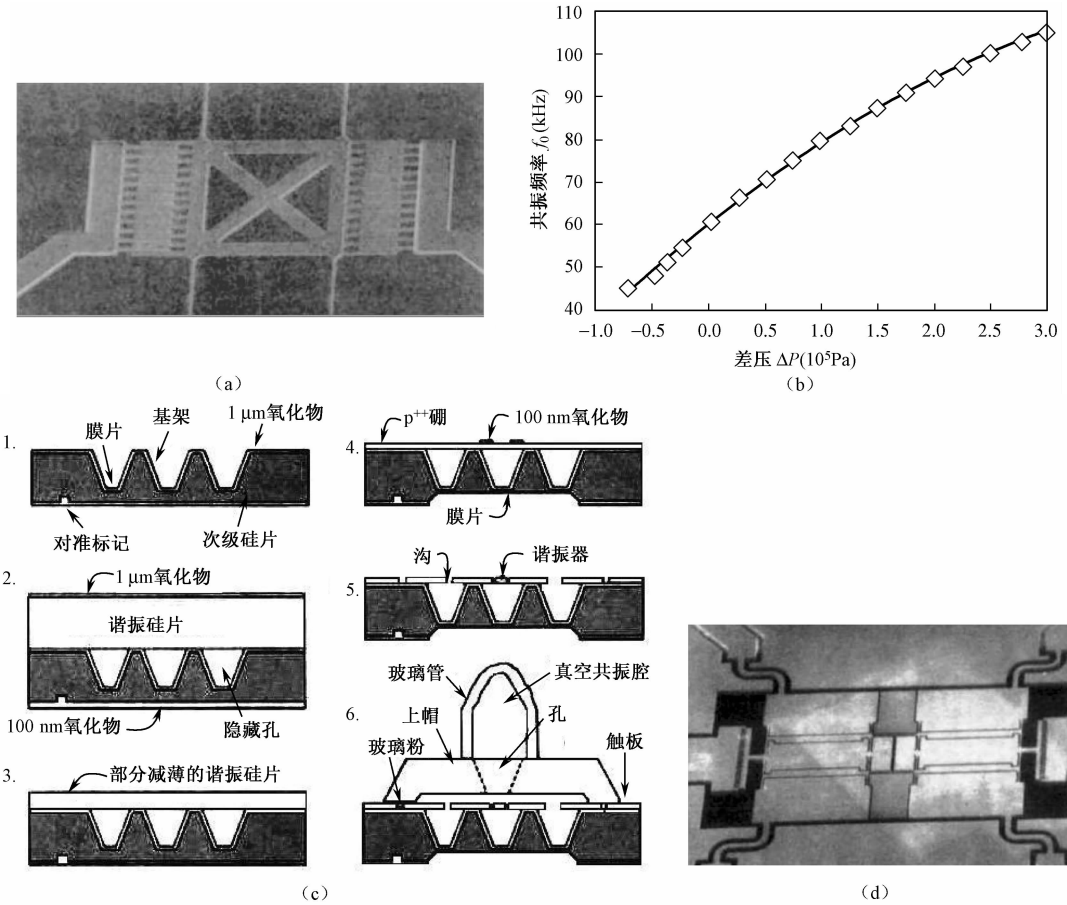


图 5.3.4 (a)以体加工工艺制造的应变元件所产生的线性振荡为基础的横向共振电容压力传感器；(b)该传感器大气压下的响应(摘自 Welham 和 coworkers. 1996)；(c)目前的硅工艺；(d)最新压阻传感器 (Welham 等, 2000)

$$m\ddot{x} + b_m\dot{x} + k_m^0 x + k_m^1 x^3 = F_x(t) \quad (5.3.4)$$

不过,因为成本低,稳定性好,电路易于集成,硅材料体加工工艺已成为当今用来生产压阻式压力传感器的首选微加工技术。表 5.3.3 总结了目前压力传感器在汽车行业的应用(madou,1997)。

显然,压力传感器的汽车市场是巨大的。目前, Motorola、Nova&Sensor、SSI Technologies 以及其他制造商已能提供大量商品化的传感器。由于成本的降低,压阻式多晶硅变得更加符合需求,但随之也存在一些稳定性和精确性方面的问题。因此,我们期待着制造膜片的替代技术例如绝缘硅(SOI)技术的出现。

表 5.3.3 1997 年硅压力传感器的应用。摘自 Madou(1997)

应 用	成 本 (欧 洲)	市 场 (中 欧)	压力范围 (kPa)	推出的年份
进气管压力	9	30	0~105	最近
大气压	9	100	50~105	最近
排气再循环	9	3.3	0~105	1989
燃料压力	9	97	0~105	1994
轮胎气压	n/a	455	500	1994—1995
液压主动悬架	7	14	20 000	1994—1995
气候控制	9	19	50~105	最近

5.3.5 Reading Materials

A pressure sensor measures the pressure, typically of gases or liquids. Pressure is an expression of the force required to stop a gas or fluid from expanding, and is usually stated in terms of force per unit area. A pressure sensor generates a signal related to the pressure imposed. Typically, such a signal is electrical, but it might also include additional means, such as optic signals, visual signals and/or auditory signals.

Pressure sensors are used in numerous ways for control and monitoring in thousands of everyday applications. Pressure sensors can be used in systems to measure other variables such as fluid/gas flow, speed, water level, and altitude. Pressure sensors can alternatively be called pressure transducers, pressure transmitters (压力变送器), pressure senders and pressure indicators(压力计), among other names.

Pressure sensors can vary drastically in technology, design, performance, application suitability and cost. A conservative estimate would be that there may be over 50 technologies and at least 300 companies making pressure sensors worldwide.

There are also a category of pressure sensors that are designed to measure in a dynamic mode for capturing very high speed changes in pressure. Example applications for this type of sensor would be in the measuring of combustion pressure (燃烧压力) in an engine cylinder or in a gas turbine. These sensors are commonly manufactured out of piezoelectric materials like quartz.

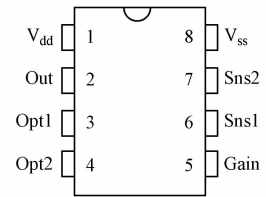
Some pressure sensors function in a binary manner, i. e. , when pressure is applied to a pressure sensor, the sensor acts to complete or break an electrical circuit. Some speed cameras use them. These types of sensors are also known as a pressure switches.

5.4 QProx™ QT113 Charge-Transfer Touch Sensor (I)

5.4.1 Text

- Projects a proximity field through air or any insulator

- Less expensive than many mechanical switches
- Sensitivity easily adjusted
- Consensus filter for noise immunity
- 100% autocal for life - no adjustments required
- 2.5 to 5 V, 600 μ A single supply operation
- Toggle mode for on/off control (strap option)
- 10 s, 60 s, infinite auto-recal timeouts (strap options)
- HeartBeat™ health indicator on output
- Only one external part required - a 1 C capacitor
- Lead-Free package



Applications

- Light switches
- Appliance control
- Access systems
- Elevator buttons
- Prox sensors
- Security systems
- Pointing devices
- Consumer devices

The QT113 charge-transfer (QT) touch sensor is a self-contained digital IC capable of detecting near-proximity or touch. It will project a proximity sense field through air, and any dielectric like glass, plastic, stone, ceramic, and most kinds of wood. It can also turn small metal-bearing objects into intrinsic sensors, making them responsive to proximity or touch. This capability coupled with its ability to self calibrate continuously can lead to entirely new product concepts.

It is designed specifically for human interfaces, like control panels, appliances, toys, lighting controls, or anywhere a mechanical switch or button may be found; it may also be used for some material sensing and control applications provided that the presence duration of objects does not exceed the recalibration timeout interval.

Power consumption is only 600 μ A in most applications. In most cases the power supply need only be minimally regulated, for example by Zener diodes or an inexpensive 3-terminal regulator. The QT113 requires only a common inexpensive capacitor in order to function.

The QT113's RISC core employs signal processing techniques pioneered by Quantum; these are specifically designed to make the device survive real-world challenges, such as 'stuck sensor' conditions and signal drift.

The option-selectable toggle mode permits on/off touch control, for example for light switch replacement. The quantum-pioneered HeartBeat™^[1] signal is also included, allowing a microcontroller to monitor the health of the QT113 continuously if desired. By using the charge transfer principle, the IC delivers a level of performance clearly superior to older technologies in a highly cost-effective package.

Available Options

TA	SOIC	8-PIN DIP
0°C to +70°C	—	QT113-DG
40°C to +85°C	QT113-ISG	—

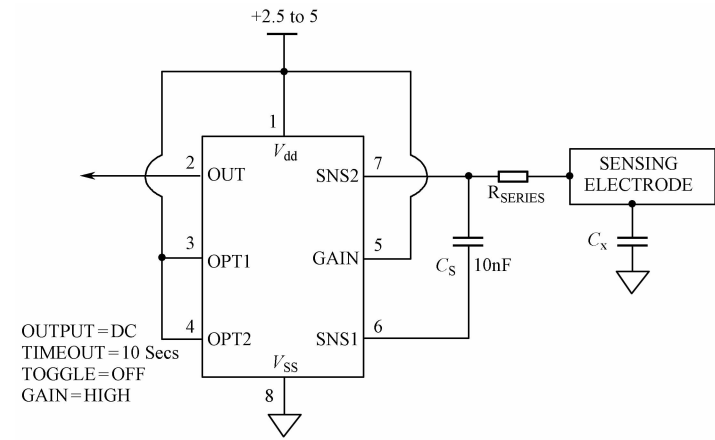


Figure 5. 4. 1 Basic Circuit Configuration.

Figure 5. 4. 1 shows a basic circuit using the device.

Basic Operation

The QT113 employs bursts of charge-transfer cycles to acquire its signal. Burst mode permits power consumption in the microamp range, dramatically reduces RF emissions, lowers susceptibility to EMI, and yet permits excellent response time. Internally the signals are digitally processed to reject impulse noise, using a ‘consensus’ filter which requires three consecutive confirmations of a detection before the output is activated.

The QT switches and charge measurement hardware functions are all internal to the QT113 (Figure 5. 4. 2). A 14-bit single-slope switched capacitor ADC includes both the required QT charge and transfer switches in a configuration that provides direct ADC conversion. The ADC is designed to dynamically optimize the QT burst length according to the rate of charge buildup on C_s , which in turn depends on the values of C_s , C_x , and V_{dd} . V_{dd} is used as the charge reference voltage. Larger values of C_x cause the charge transferred into C_s to rise more rapidly, reducing available resolution; as a minimum resolution is required for proper operation, this can result in dramatically reduced apparent gain. Conversely, larger values of C_s reduce the rise of differential voltage across it, increasing available resolution by permitting longer QT bursts. The value of

Overview

The QT113 is a digital burst mode charge-transfer (QT) sensor designed specifically for touch controls; it includes all hardware and signal processing functions necessary to provide stable sensing under a wide variety of changing conditions. Only a single low cost, noncritical capacitor is required for operation.

C_s can thus be increased to allow larger values of C_x to be tolerated.

The IC is responsive to both C_x and C_s , and changes in C_s can result in substantial changes in sensor gain.

Option pins allow the selection or alteration of several special features and sensitivity.

Electrode Drive

The internal ADC treats C_s as a floating transfer capacitor; as a result, the sense electrode can in theory be connected to either SNS1 or SNS2 with no performance difference.

However the electrode should only be connected to pin SNS2 for optimum noise immunity.

In all cases the rule $C_s \gg C_x$ must be observed for proper operation; a typical load capacitance (C_x) ranges from 10-20 pF while C_s is usually around 10-50 nF.

Increasing amounts of C_x destroy gain; therefore it is important to limit the amount of stray capacitance on both SNS terminals, for example by minimizing trace lengths and widths and keeping these traces away from power or ground traces or copper pours.

The traces and any components associated with SNS1 and SNS2 will become touch sensitive and should be treated with caution to limit the touch area to the desired location.

A series resistor, R_{series} , should be placed inline with the SNS2 pin to the electrode to suppress ESD and EMC effects.

Electrode Design

Electrode Geometry and Size. There is no restriction on the shape of the electrode; in most cases common sense and a little experimentation can result in a good electrode design. The QT113 will operate equally well with long, thin electrodes as with round or square ones; even random shapes are acceptable. The electrode can also be a 3-dimensional surface or object. Sensitivity is related to electrode surface area, orientation with respect to the object being sensed, object composition, and the ground coupling quality of both the sensor circuit and the sensed object^[2].

If a relatively large electrode surface is desired, and if tests show that the electrode has more capacitance than the QT113 can tolerate, the electrode can be made into a sparse mesh (Figure 5. 4. 3) having lower C_x than a solid plane. Sensitivity may even remain the same, as the sensor will be operating in a lower region of the gain curves.

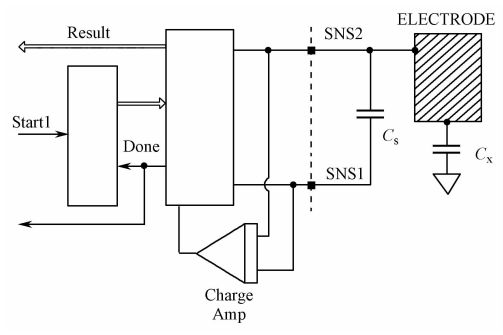


Figure 5. 4. 2 Internal Switching & Timing.

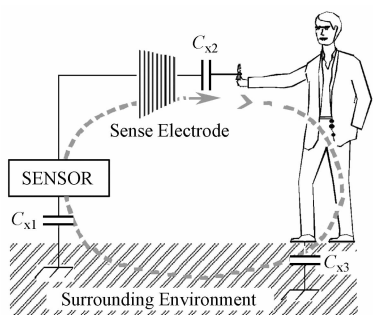


Figure 5. 4. 3 Kirchoff's Current Law.

Kirchoff's Current Law. Like all capacitance sensors, the QT113 relies on Kirchoff's Current Law (Figure 5. 4. 3) to detect the change in capacitance of the electrode. This law as applied to capacitive sensing requires that the sensor's field current must complete a loop, returning back to its source in order for capacitance to be sensed^[3]. Although most designers relate to Kirchoff's law with regard to

hardwired circuits, it applies equally to capacitive field flows. By implication it requires that the signal ground and the target object must both be coupled together in some manner for a capacitive sensor to operate properly. Note that there is no need to provide actual hardwired ground connections; capacitive coupling to ground (C_{x1}) is always sufficient, even if the coupling might seem very tenuous. For example, powering the sensor via an isolated transformer will provide ample ground coupling, since there is capacitance between the windings and/or the transformer core, and from the power wiring itself directly to 'local earth'. Even when battery powered, just the physical size of the PCB and the object into which the electronics is embedded will generally be enough to couple a few picofarads back to local earth.

5. 4. 2 Specialized English Words

charge-transfer(QT) 电荷转移
project 投射
proximity field 接近场
consensus filter 协同滤波器
noise immunity 抗(干)扰性
autocal = auto calibrate 自动校正,
自动标定
toggle mode 切换方式
strap option 搭接选择,接线选择
lead-free package 无铅封装
appliance control 家电控制
access systems 接入系统,接入技术
prox sensors 接近传感器,接近开关
pointing device 指点装置,定位装置

consume devices 消费设备
self-contained 独立的,自成体系
dielectric 绝缘的,(电)介质
ceramic 陶瓷
metal-bearing 含金属的
burst 脉冲(串)
electrode 电极
TA 即 TAPP,纤薄阵列塑料封装
(Thin Array Plastic Package之缩写)
SOIC 小输出线封装(Small Outline
IC之缩写)
DIP 双列直插式封装(Dual In-line
Package之缩写)
non-critical 不挑剔的,要求不高的

RF emissions 射频辐射(RF 为 radio frequency 之缩写)	ESD 静电放电(Electrostatic discharge 之缩写)
susceptibility 敏感性,可感性	sparse 稀疏的
EMI 电磁干扰(Electro Magnetic Interference 之缩写)	mesh 网(状物)
consecutive 一致的	by implication 其实
single-slope 单斜坡	tenuous 脆弱的,微弱的
ADC 模数转换(部件)(Analog Digital Conversion 之缩写)	ample 足够的,丰裕的
stray capacitance 杂散电容,分布电容	PCB 印制电路板(Printed Circuit board 之缩写)
trace 引线	wind 绕组
copper pours (印制电路板上)敷铜	core (铁)芯
	picofarad 微微法(拉),皮(法拉)

5.4.3 Notes

[1] HeartBeatTM. 这里 TM 为 Trade Mark 之缩写,“注册商标”之意,表示 HeartBeatTM是一注册商标,受产权保护。

[2] Sensitivity is related to electrode surface area, orientation with respect to the object being sensed, object composition, and the ground coupling quality of both the sensor circuit and the sensed object. 这是一个简单句,“is related to”后有多介词宾语。注意其中“the object being sensed”和“the sensed object”,前者为现在分词短语做后置定语,后者为过去分词做前置定语,意思完全相同。全句可译为“灵敏度和电极表面面积、感应对象的朝向、材质以及感应电路和感应对象两者对地的耦合质量等有关。”

[3] This law as applied to capacitive sensing requires that the sensor's field current must complete a loop, returning back to its source in order for capacitance to be sensed. 这个主从复合句中,“as applied to capacitive sensing”是主句的状语,可视为状语从句“as it is applied…”的省略。注意“that”引起的宾语从句中“in order to”后接的不定式动词原形因为逻辑主语“(for) capacitor”而变成被动语态“(to) be sensed”了。全句可译为“用于电容传感时,该定律要求传感器的场电流必须构成一个回路,最后返回源头,才能感应到被测电容。”

5.4.4 Reference Translation

电荷转移触摸传感器 QProxTM QT113(I)

- 建立一个穿透空气和任何绝缘物体的接近场
- 比众多机械开关价格低廉
- 灵敏度调整简单
- 抗干扰协同滤波

- 100%终身自动校正——无需人工调整
- 2.5~5 V, 600 μ A 单电源运行
- 脉冲输出选择方式(搭接选择)
- 10 秒, 60 秒, 无限长超时自动校正(搭接选择)
- HeartbeatTM输出正常指示器
- 仅需一个外接元件——一个价值一美分的电容
- 无铅封装

应用

- 灯具开关
 - 接近传感器
- 家电控制
 - 安防系统
- 接入系统
 - 指点装置
- 电梯按钮
 - 消费设备

QT113 电荷转移(QT)式触摸传感器是一个自成体系的数字集成电路芯片,能对接近信号和触摸信号进行检测。QT113 建立起一个接近感应场,可以穿过空气及任何绝缘材料,如玻璃、塑料、石头、陶瓷及大多数木材。它可以将含金属的小物体转变为像传感器一样,可以对接近信号和触摸信号做出反应。这一性能与它的连续自动校正功能相结合,产生出传感器的一种全新的概念。QT113 是专门为人机接口设计的,例如控制面板、家用电器、玩具、照明控制,或者说任何机械开关和按钮;它还可以用于某些材料的探测和控制应用,只要这时对象作用的时间不超过再校正超时时间间隔即可。

大多应用场合的电耗仅 600 μ A。多数情况下,电源只需简单稳压即可,例如采用齐纳二极管或廉价的三端稳压电源。QT113 仅需接上一个普通的廉价电容便可工作。

QT113 的 RISC 内核使用了 Quantum 首创的多种信号处理技术,使 QT113 面临实际应用环境的挑战,如“传感阻塞”、信号漂移时也能正常工作。

可选用的脉动方式能进行通/断控制,例如代替灯具开关。QT113 还有 Quantum 首创的 HeartBeatTM信号输出,使得外接的微控制器可以根据需要连续监视 QT113 工作是否正常。由于采用了电荷转移原理,QT113 性能上明显优于老技术且成本十分低廉。

可供品种

TA 封装	SOIC 封装	8 脚 DIP 封装
0℃~+70℃	—	QT113-DG
40℃~+85℃	QT113-ISG	—

综述

QT113 是专为触摸控制而设计的数字脉冲式电荷转移(QT)传感器,它包含了所有必要的硬件和信号处理功能,可以在各种变化的环境下提供稳定的信号传感。只需配接一个并无特殊要求的廉价电容即可工作。图 5.4.1 所示为使用 QT113 的基本电路图。

基本运行状况

QT113 通过电荷转移周期脉冲串来获取信号,这种脉冲串方式可以把电耗控制在微安级,这大大减少了射频辐射,降低了电磁干扰敏感度,而且能得到极佳的响应时间。在 QT113 内部,采用了“协同”滤波器进行数字抗脉冲干扰处理。对每次检测,都要求得到

三次测量的一致确认后才对外输出。

QT 将电荷转换和测量的全部硬件功能部件都已纳入芯片之中（见图 5.4.2）。14 位单斜坡开关电容模数转换器同一个部件可施行 QT 所需的充电和转换两者间的切换控制，直接进行模数转换。ADC 可根据 C_s 上电荷充电速度动态地优化 QT 脉冲串数量。充电速度与 C_s 、 C_x 和 V_{dd} 的值有关。 V_{dd} 用

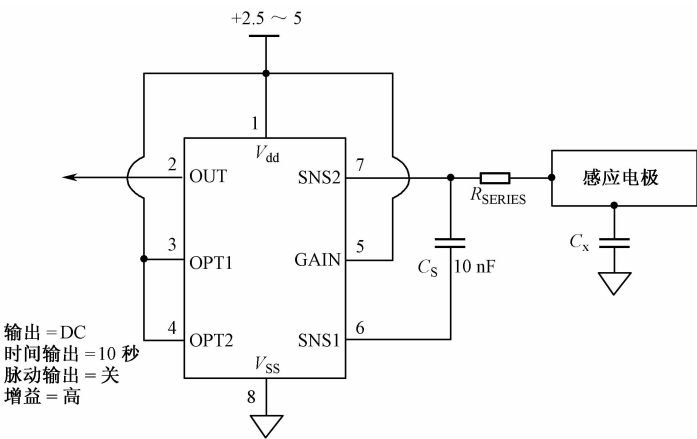


图 5.4.1 基本电路组成

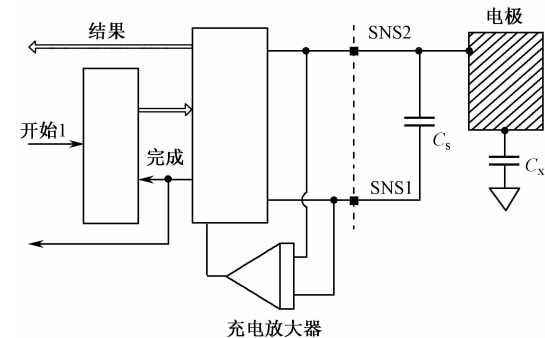


图 5.4.2 内部转换与定时

电极驱动

内部 ADC 将 C_s 视为一个浮动的转换电容。这样，理论上感应电极就可以接到 SNS1 或 SNS2 上，两种接法性能上没有差别。然而，只有将电极与 SNS2 相连才能将干扰降到最低。

任何情况下必须保证 $C_s \gg C_x$ 才能正常运行。典型负载电容 (C_x) 之值为 10~20 pF，而 C_s 通常为 10~50 nF。

增加 C_x 之值将恶化增益值。所以减小两个 SNS 端口的杂散电容是很重要的。例如，尽量减小引线的长度和宽度，让引线远离电源、地线以及印制电路板的敷铜区。

引线及任何与 SNS1、SNS2 相连的元件都和触摸灵敏度有关，必须小心处置，使触摸区能处于所需要的位置上。

串联电阻 R_{series} 应串接在 SNS2 引脚和电极之间，用以降低静电放电和电磁干扰的影响。

电极设计

电极尺寸与大小。对电极形状无任何限制。一般情况下，只要简单地实际触摸试验一下就可以得到一个良好的电极设计。无论是长的薄的电极，还是圆的方的电极，QT113 运行一样良好。其实任意形状都是可以接受的。电极还可以是一个三维的表面或物体。灵敏度和电极表面面积、感应对象的朝向、材质以及感应电路和感应对象两者对地的耦合质量等有关。

如果希望采用大面积的电极，而且测试表明电极电容比 QT113 所要求的大，这时可将电极制成稀疏的网状物，它的 C_x 比一个整体面的电容量要小。这样灵敏度可能依然不变，因为 QT113 将运行在增益曲线的较低端。

基尔霍夫电流定律。和所有电容传感器一样，QT113 根据基尔霍夫电流定律(见图

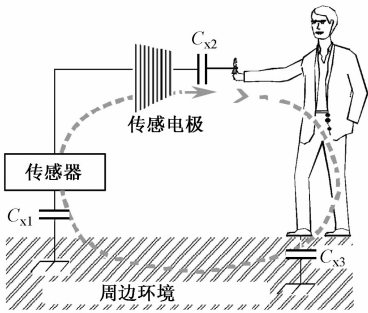


图 5.4.3 基尔霍夫电流定律

5.4.3)检测电极电容的变化。用于电容传感时，该定律要求传感器的场电流必须构成一个回路，最后返回源头，才能感应到被测电容。尽管多数设计者将基尔霍夫定律和硬接线电路联系到一起，其实基尔霍夫定律同样适用于电容场电流。实际上，电容传感器要能正常工作的话，它要求信号地和感应目标必须以某种方式联系起来。注意，这里无需实际的地线连接。尽管看上去接地很弱，但电容对地的连接(C_{x1})总是很充分的。例如，向传感器供电的隔离变压器也可以提供良好的接地耦合，因为变压器绕组和铁心之间存在电容，又经电源线直接和“本地”地相连。甚至在电池供电时，就是印制电路板板体和传感器所嵌入其中的设备一般也可以对地耦合出几个微微法的电容。

5.4.5 Reading Materials

Output Features of QT113

The QT113 is designed for maximum flexibility and can accommodate most popular sensing requirements. These are selectable using strap options on pins OPT1 and OPT2. All options are shown in Table. 5.4.1.

DC Mode Output

The output of the QT113 can respond in a DC mode, where the output is active-low (低电平有效) upon detection. The output will remain active-low for the duration of the detection, or until the Max On-Duration(最大输出持续时间) expires (if not infinite), whichever occurs first. If a max on-duration timeout occurs first, the sensor performs a full recalibration and the output becomes inactive until the next detection.

In this mode, three Max On-Duration timeouts are available: 10 seconds, 60 seconds, and infinite. Infinite timeout is useful in applications where a prolonged(长时

间的) detection can occur and where the output must reflect the detection no matter how long.

In infinite timeout mode, the designer should take care to be sure that drift in C_s , C_x , and V_{dd} do not cause the device to ‘stick on’ inadvertently even when the target object is removed from the sense field.

Toggle Mode Output (脉动输出方式)

This makes the sensor respond in an on/off mode(通断方式) like a flip flop. It is most useful for controlling power loads, for example in kitchen appliances, power tools, light switches, etc.

Max On-Duration in Toggle mode is fixed at 10 seconds. When a timeout(超时) occurs, the sensor recalibrates(重新标定) but leaves the output toggle state unchanged.

Heartbeat™ Output

The QT113 output has a full-time HeartBeat™ ‘health’ indicator superimposed on it. This operates by taking ‘Out’ into a 3-state mode for 300 μs once after every QT burst. This output state can be used to determine that the sensor is operating properly, or, it can be ignored using one of several simple methods. The HeartBeat indicator can be sampled by using a pulldown resistor on Out, and feeding the resulting negative-going pulse(负脉冲) into a counter, flip flop, one-shot(单脉冲), or other circuit. Since Out is normally high, a pulldown resistor will create negative HeartBeat pulses when the sensor is not detecting an object; when detecting an object, the output will remain low for the duration of the detection, and no HeartBeat pulse will be evident.

5.5 QProx™ QT113 Charge-Transfer Touch Sensor (II)

5.5.1 Text

Virtual Capacitive Grounds

When detecting human contact (e. g. a fingertip), grounding of the person is never required. The human body naturally has several hundred picofarads of ‘free space’ capacitance to the local environment (C_{x3} in Figure 5. 4. 3), which is more than two orders of magnitude greater than that required to create a return path to the QT113 via earth^[1]. The QT113’s PCB however can be physically quite small, so there may be little ‘free space’ coupling (C_{x1} in Figure 5. 4. 3) between it and the environment to complete the return path. If the QT113 circuit ground cannot be earth grounded by wire, for example via the supply connections, then a ‘virtual capacitive ground’ may be required to increase return coupling.

A ‘virtual capacitive ground’ can be created by connecting the QT113’s own circuit

ground to:

- A nearby piece of metal or metalized housing
- A floating conductive ground plane
- Another electronic device (to which its output might be connected anyway)

Free-floating ground planes such as metal foils should maximize exposed surface area in a flat plane if possible. A square of metal foil will have little effect if it is rolled up or crumpled into a ball. Virtual ground planes are more effective and can be made smaller if they are physically bonded to other surfaces, for example a wall or floor.

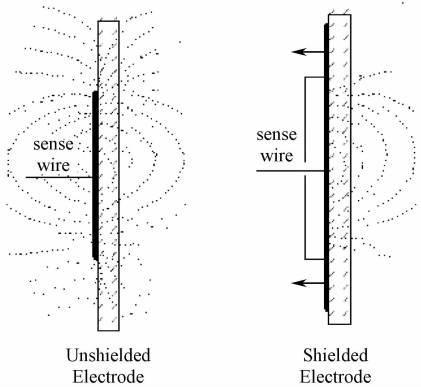


Figure 5.5.1 Shielding Against Fringe Fields.

electrode size and capacitance, electrode shape and orientation, the composition and aspect of the object to be sensed, the thickness and composition of any overlaying panel material, and the degree of ground coupling of both sensor and object.

Table 5.5.1 Gain Setting Strap Options.

Gain	Tie Pin 5 to:
High — 6 counts	V_{dd}
Low — 12 counts	$V_{ss}(\text{Gnd})$

Increasing Sensitivity

In some cases it may be desirable to increase sensitivity further, for example when using the sensor with very thick panels having a low dielectric constant^[2].

Sensitivity can often be increased by using a bigger electrode, reducing panel thickness, or altering panel composition. Increasing electrode size can have diminishing returns, as high values of C_x will reduce sensor gain. The value of C_s also has a dramatic effect on sensitivity, and this can be increased in value with the tradeoff of reduced response time. Increasing the electrode’s surface area will not substantially increase touch sensitivity if its diameter is already much larger in surface area than the object

being detected^[3]. Panel material can also be changed to one having a higher dielectric constant, which will help propagate the field. Metal areas near the electrode will reduce the field strength and increase C_x loading.

Ground planes around and under the electrode and its SNS trace will cause high C_x loading and destroy gain. The possible signal-to-noise ratio benefits of ground area are more than negated by the decreased gain from the circuit, and so ground areas around electrodes are discouraged. Keep ground away from the electrodes and traces.

QT113 Specifics

Signal Processing

The QT113 processes all signals using 16 bit math, using a number of algorithms pioneered by Quantum. The algorithms are specifically designed to provide for high ‘survivability’ in the face of numerous adverse environmental changes.

Drift Compensation Algorithm

Signal drift can occur because of changes in C_x and C_s over time. It is crucial that drift be compensated for, otherwise false detections, non-detections, and sensitivity shifts will follow.

Drift compensation (Figure 5. 5. 2) is performed by making the reference level track the raw signal at a slow rate, but only while there is no detection in effect. The rate of adjustment must be performed slowly, otherwise legitimate detections could be ignored. The QT113 drift compensates using a slew-rate limited change to the reference level; the threshold and hysteresis values are slaved to this reference.

Once an object is sensed, the drift compensation mechanism ceases since the signal is legitimately high, and therefore should not cause the reference level to change.

The QT113’s drift compensation is ‘asymmetric’: the reference level drift-compensates in one direction faster than it does in the other.

Specifically, it compensates faster for decreasing signals than for increasing signals. Increasing signals should not be

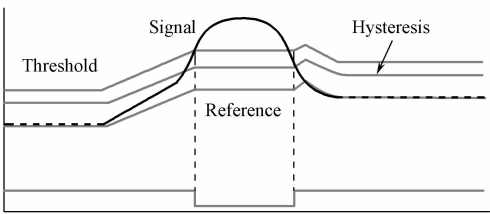


Figure 5. 5. 2 Drift Compensation.

compensated for quickly, since an approaching finger could be since an approaching finger could be compensated for partially or entirely before even approaching the sense electrode. However, an obstruction over the sense pad, for which the sensor has already made full allowance for, could suddenly be removed leaving the sensor with an artificially elevated reference level and thus become insensitive to touch. In this latter case, the sensor will compensate for the object’s removal very quickly, usually in only a

few seconds.

With large values of C_s and small values of C_x , drift compensation will appear to operate more slowly than with the converse. Note that the positive and negative drift compensation rates are different.

Threshold Calculation

The internal threshold level is fixed at one of two setting as determined by Table 5. 5. 1. These settings are fixed with respect to the internal reference level, which in turn will move in accordance with the drift compensation mechanism.

The QT113 employs a hysteresis dropout below the threshold level of 17% of the delta between the reference and threshold levels.

Max On-duration

If an object or material obstructs the sense pad the signal may rise enough to create a detection, preventing further operation. To prevent this, the sensor includes a timer which monitors detections. If a detection exceeds the timer setting, the timer causes the sensor to perform a full recalibration (when not set to infinite). This is known as the Max On-Duration feature.

After the Max On-Duration interval, the sensor will once again function normally to the best of its ability given electrode conditions. There are two finite timeout durations available via strap option: 10 and 60 seconds (Table 5. 5. 2).

Table 5.5.2 Output Mode Strap Options.

	Tie Pin 3 to	Tie Pin 4 to	Max On- Duration
DC Out	V_{dd}	V_{dd}	10 s
DC Out	V_{dd}	Gnd	60 s
Toggle	Gnd	Gnd	10 s
DC Out	Gnd	V_{dd}	Infinite

Detection Integrator

It is desirable to suppress detections generated by electrical noise or from quick brushes with an object. To accomplish this, the QT113 incorporates a detect integration counter that increments with each detection until a limit is reached, after which the output is activated^[4]. If no detection is sensed prior to the final count, the counter is reset immediately to zero. In the QT113, the required count is 3.

The Detection Integrator can also be viewed as a ‘consensus’ filter, that requires three successive detections to create an output.

5. 5. 2 Specialized English Words

metalized housing (镀)金属外壳 crumple into 弄皱

fringe fields 散射(电)场,干扰(电)场
lateral 横向的,侧面的
inadvertent 无意的,偶然的
diminishing returns 得不偿失
trade-off 权衡,折中
propagate 传播,繁殖
adverse 不利的,有害的
survivability 生存能力
legitimate 合理的,正当的,合法的

slew-rate 转换斜率,边沿斜率
threshold 门槛,阈值,临界值
hysteresis 滞后,延滞
asymmetric 不对称的
obstruction 堵塞(物),阻挡(物)
elevated 升高的,高贵的
suppress 抑制,排除,镇压
brushes 轻擦,擦过
successive 连续的,逐次的

5.5.3 Notes

[1] The human body naturally has several hundred picofarads of ‘free space’ capacitance to the local environment (C_{x3} in Figure 5.4.3), which is more than two orders of magnitude greater than that required to create a return path to the QT113 via earth. 在这个复合句中,“which”引导的定语从句修饰“several hundred picofarads”,从句中的“that”为代词,其后的“required to create a return path to the QT113 via earth”是它的定语。全句译为“人体自然而然地会和周边环境产生出一个几百个微微法的‘自由空间’电容(图 5.4.3 中的 C_{x3}),这比 QT113 建立经过大地的回路所需要的电容值要大两个数量级。”

[2] In some cases it may be desirable to increase sensitivity further, for example when using the sensor with very thick panels having a low dielectric constant. 这是一个简单句。注意“when using ...”可以认为是“when you are using the sensor with very thick panels having a low dielectric constant”省去了“you are”。全句可译为“某些情况下,希望将灵敏度进一步增加,例如在传感器的面板很厚而介电常数又很小的情况下。”

[3] Increasing the electrode’s surface area will not substantially increase touch sensitivity if its diameter is already much larger in surface area than the object being detected. 在这个复合句中,“Increasing the electrode’s surface area”为动名词短语做主句的主语。“if its diameter is already much larger in surface area than the object being detected.”为条件从句,其中“being detected”为被动语态的现在分词,做“object”的后置定语。全句可译为“在电极面积已经比被测物的面积大很多的情况下,如果再增大电极面积并不会明显提高触摸灵敏度。”

[4] To accomplish this, the QT113 incorporates a detect integration counter that increments with each detection until a limit is reached, after which the output is activated. 在这个复合句中,主句为“the QT113 incorporates a detect integration counter”,“that increments with each detection until a limit is reached”为“a detect integration counter”的定语从句,“until a limit is reached”则是从句的状语从句,而“after which the output is activated”则是与“that increments with each detection”并列的从句,表示进一步的结果。由以上分析可以看到本句共有四个句子,结构上是较复杂的。全句

可译为“为了实现这一目标,QT113 整合了一个检测计数器。这个计数器在每次检测时就开始增量,到达上限值时,才激活输出。”

5.5.4 Reference Translation

电荷转移触摸传感器 QProx™ QT113(II)

虚拟电容地

在检测人体触摸信号(即手指)时,完全不需要人体接地。人体自然而然地会和周边环境产生出一个几百微微法的“自由空间”电容(图 5.4.3 中的 C_{x3}),这比 QT113 建立经过大地回路所需的值要大两个数量级。但是 QT113 的电路板有可能实际尺寸太小,导致它和周边环境之间的“自由空间”耦合(图 5.4.3 中的 C_{x1})太弱,不足以形成回路。如果 QT113 电路的地无法通过导线,例如电源接地,那么可能需要“虚拟电容地”来加强回路耦合。

- “虚拟电容地”可以通过将 QT113 本身的地做下述连接而接地：
- 附近的一块金属体或者金属壳体
 - 浮置的导电地平面
 - 另一电气设备(其实 QT113 的输出可能就是与它相连的)

自由浮动的接地面,如金属箱,应当尽可能地暴露面展开成平面。如果将一块方形金属箱卷起来或者揉成球,效果就很差。虚拟地表如果能和其他表面如墙体或地板贴附在一起,则效果更明显,而且可做得小一些。

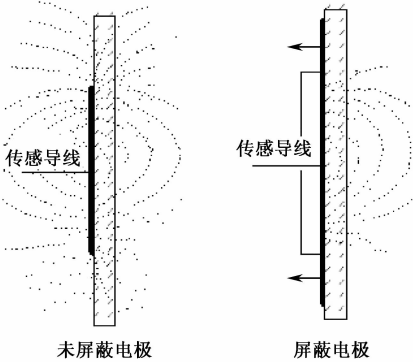


图 5.5.1 散射(电)场

电场的形成

图 5.5.1 给出了散射(电)场的屏蔽……

灵敏度

通过选接引脚 P5(见表 5.5.1),QT113 可以从两个增益中选取一个进行设置。这样改变灵敏度是通过内部变更检测所需的一个数字门槛来实现的。注意,灵敏度也和其他一些因素都有关,如 C_s 的值,电极的大小、电容量、形状、朝向,被测物的成分和形状,包裹电极板的材料成分,以及传感器和被测物对地的耦合程度。

表 5.5.1 增益设定接线

增 益	引脚 5 接至
高 — 6 计数	V_{dd}
低 — 12 计数	$V_{ss}(\text{Gnd})$

增加灵敏度

某些情况下,希望将灵敏度进一步增加,例如在传感器的面板很厚而介电常数又很小的情况下。

通常可通过增大电极面积、减小面板厚度或者改变面板材质的办法提高灵敏度。增大电极面积的做法可能得不偿失,因为 C_x 值增大会降低传感器增益。 C_s 值也会对灵敏度有明显影响,而 C_s 值增大的负面影响是恶化了响应时间。在电极面积已经比被测物的面积大很多的情况下,如果再增大电极面积并不会明显提高触摸灵敏度。面板材料可改用较高介电常数的材料,以增强电场的辐射。电极附近的金属体将会降低电场强度,增大 C_x 负载。

电极及 SNS 接线周围或下方的地面也会增大 C_x 负载,减小增益。这样可能带来的信噪比方面的好处被电路增益减少带来的不利大大抵消。所以不宜让电极周围地面积过大,让电极和接线远离地面。

QT113 特性

信号处理

QT113 采用 Quantum 开发的一系列算法对信号进行 16 位数字处理。这些算法都是专门设计的,使 QT113 在各种不利的环境变化下仍能正常运行。

漂移补偿算法

C_x 和 C_s 随时间而变化时会发生信号漂移。对漂移进行补偿至关重要,否则,误测、漏测、灵敏度偏移等问题都会随之而来。

漂移补偿(见图 5.5.2)是通过让参考电位以一种慢速率方式跟随原始信号来实现的,但只有在检测未进行时才进行补偿。

补偿调整的速度必须很慢,否则正常的检测可能会被漏过。QT113 用一种缓慢的变化率对参考电压进行补偿,阈值和滞后值则跟随参考值而变化。一旦测到触摸信号,漂移补偿机制便停止工作,因为这时信号的升高是正当的,不应引起参考电压的改变。QT113 对参考电压的漂移补偿是“非对称的”,即在一个方向比在另一个方向上更快。具体而言,对信号减小时的补偿比对信号增大时的补偿来得快些。之所以对增大的信号不应补偿过快,是因为一个正在按下的手指可以在手指未按到感应电极之前就可对其进行部分或全部的补偿。然而,感应电极之上的手指,有可能突然移开(对此传感器已经做了最大的允许设定),结果使得传感器呈现出一个人工升高的参考值,造成传感灵敏度下降。对后面这一情况,传感器一般会在手指移去后仅几秒钟之内快速做出补偿。

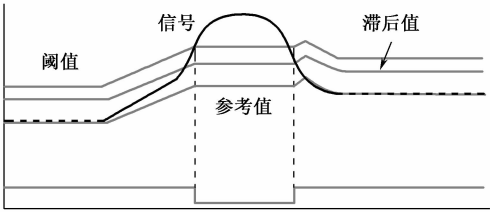


图 5.5.2 漂移补偿

C_s 大、 C_x 小时的漂移补偿比 C_s 小、 C_x 大时的漂移补偿显得更慢。注意,正补偿和负补偿的速度是不一样的。

阈值的标定

内部阈值可根据表 5.5.1 的选择确定为两个设定值之一,值的最后确定还与内部参考电压值有关,而参考电压又随漂移补偿机制的作用而移动。

QT113 在参考电压和阈值电压之间有滞后电压,其值比参考电压低 17%。

最大输出时间

感应物体作用在感应电极上,信号就会升得足够高,从而进入一次检测,并阻止进一步的操作。为了防止这种情况发生,QT113 有一个定时器对此进行监视。如果某次检测超过了定时器设定值,定时器就会引发传感器进行完全的重新标示(不设定在无限长时),这称为最大定时特性。

超过最大定时间隔后,传感器将重新按照最佳功能设置电极条件。通过引脚选择连线可以选用两个限超时间隔:10 秒或 60 秒(见表 5.5.2)。

表 5.5.2 输出模式接线选择

	引脚 3 接至	引脚 4 接至	最大接通时间
DC 输出	V_{dd}	V_{dd}	10 s
DC 输出	V_{dd}	Gnd	60 s
脉动输出	Gnd	Gnd	10 s
DC 输出	Gnd	V_{dd}	无限长

检测计数器

对于电气噪声和感应对象的划擦产生的感应信号应加以抑制。为了实现这一目标,QT113 整合了一个检测计数器。这个计数器在每次检测时就开始增量,到达上限值时,才激活输出。如果在记数上限值之前未测到感应信号,计数器立即复位清零。QT113 的计数值为 3。

这个检测计数器也可视为一个“一致性”滤波器,必须连续 3 次都测到感应值才有输出。

5.5.5 Reading Materials

Power Supply,PCB Layout of QT113

The power supply can range from 2.5 to 5.0 volts. At 3 volts current drain(电流消耗) averages less than 600 μA in most cases, but can be higher if C_s is large. Increasing C_x values will actually decrease power drain. Operation can be from batteries, but be cautious about loads causing supply droop.

As battery voltage sags(电压下降) with use or fluctuates(波动) slowly with temperature, the QT113 will track and compensate for these changes automatically with only minor changes in sensitivity.

If the power supply is shared with another electronic system, care should be taken to assure that the supply is free of digital spikes, sags, and surges(尖峰,下凹,浪涌) which can adversely affect the QT113. The QT113 will track slow changes in V_{dd} , but it

can be affected by rapid voltage steps.

If desired, the supply can be regulated using a conventional low current regulator, for example CMOS regulators that have low quiescent currents. Bear in mind that such regulators generally have very poor transient line and load stability; in some cases, shunting V_{dd} to V_{ss} with a 4.7 K resistor to induce a continuous current drain can have a very positive effect on regulator performance.

Parts placement: The chip should be placed to minimize the SNS2 trace length to reduce low frequency pickup(传感, 探测), and to reduce stray C_x which degrades gain. The C_s and R series resistors (see Figure 5.4.1) should be placed as close to the body of the chip as possible so that the SNS2 trace between Rseries and the SNS2 pin is very short, thereby reducing the antenna-like ability of this trace to pick up high frequency signals and feed them directly into the chip.

For best EMC(电磁兼容) performance the circuit should be made entirely with EMC components. **SNS trace routing:** Keep the SNS2 electrode trace (and the electrode itself) away from other signal, power, and ground traces including over or next to ground planes. Adjacent switching signals can induce noise onto the sensing signal; any adjacent trace or ground plane next to or under either SNS trace will cause an increase in C_x load and desensitize the device.

For proper operation a 100 nF (0.1 μ F) ceramic bypass capacitor must be used directly between V_{dd} and V_{ss} ; the bypass cap(旁路电容) should be placed very close to the device's power pins.

Part 6 Electric Devices and Systems

6.1 Transformers

6.1.1 Text

Description

Although transformers have no moving parts, they are essential to electromechanical energy conversion. They make it possible to increase or decrease the voltage so that power can be transmitted at a voltage level that results in low costs, and can be distributed and used safely^[1]. In addition, they can provide matching of impedances, and regulate the flow of power (real or reactive) in a network.

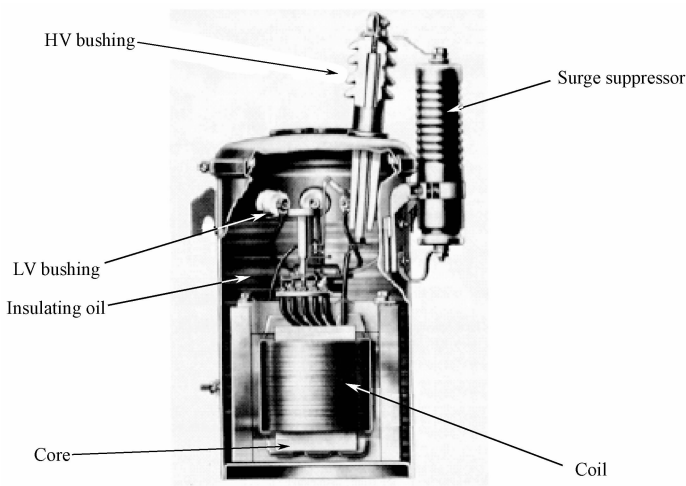


Figure 6.1.1 Cutaway view of a single phase distribution transformer. Notice only one HV bushing and lightning arrester.

When we see a transformer on a utility pole all we see is a cylinder with a few wires sticking out. These wires enter the transformer through bushings that provide isolation between the wires and the tank. Inside the tank there is an iron core linking coils, most probably made with copper, and insulated. The system

of insulation is also associated with that of cooling the core/coil assembly^[2]. Often the insulation is paper, and the whole assembly may be immersed in insulating oil, used to both increase the dielectric strength of the paper and to transfer heat from the core-coil assembly to the outer walls of the tank to the air. Figure 6.1.1 shows the cutout of a typical distribution transformer.

The Ideal Transformer

Few ideal versions of human constructions exist, and the transformer offers no exception. An ideal transformer is based on very simple concepts, and a large number of assumptions. This is the transformer one learns about in high school.

Let us take an iron core with infinite permeability and two coils wound around it (with zero resistance), one with N_1 and the other with N_2 turns, as shown in Figure 6. 1. 2. All the magnetic flux is to remain in the iron. We assign dots at one terminal of each coil in the

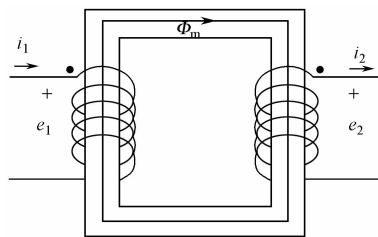


Figure 6. 1. 2 Magnetic Circuit of an ideal transformer.

following fashion; if the flux in the core changes, inducing a voltage in the coils, and the dotted terminal of one coil is positive with respect its other terminal, so is the dotted terminal of the other coil^[3]. Or, the corollary to this, current into dotted terminals produces flux in the same direction.

Assume that somehow a time varying flux, $\Phi(t)$, is established in the iron. Then the flux linkages in each coil will be $\lambda_1 = N_1 \Phi(t)$ and $\lambda_2 = N_2 \Phi(t)$. Voltages will be induced in these two coils:

$$e_1(t) = \frac{d\lambda_1}{dt} = N_1 \frac{d\Phi}{dt} \quad (6.1.1)$$

$$e_2(t) = \frac{d\lambda_2}{dt} = N_2 \frac{d\Phi}{dt} \quad (6.1.2)$$

and dividing:

$$\frac{e_1(t)}{e_2(t)} = \frac{N_1}{N_2} \quad (6.1.3)$$

On the other hand, currents flowing in the coils are related to the field intensity H . If currents flowing in the direction shown, i_1 into the dotted terminal of coil 1, and i_2 out of the dotted terminal of coil 2, then

$$N_1 \cdot i_1(t) - N_2 \cdot i_2(t) = H \cdot l \quad (6.1.4)$$

but $B = \mu_{\text{iron}} H$, and since B is finite and μ_{iron} is infinite, then $H = 0$. We recognize that this is practically impossible, but so is the existence of an ideal transformer. Finally,

$$\frac{i_1}{i_2} = \frac{N_2}{N_1} \quad (6.1.5)$$

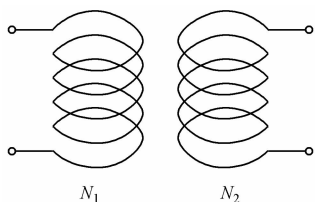


Figure 6. 1. 3 Symbol for an ideal transformer.

Equations (6. 1. 3) and (6. 1. 5) describe this ideal transformer, a two port network. The symbol of a network that is defined by these two equations is in the Figure 6. 1. 3. An ideal transformer has an interesting characteristic. A two-port network that contains it and impedances

can be replaced by an equivalent other, as discussed below. Consider the circuit in Figure 6.1.4(a). Seen as a two port network.

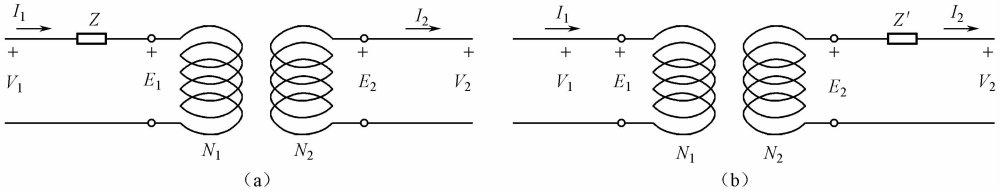


Figure 6.1.4 Transferring an impedance from one side to the other of an ideal transformer.

with variables v_1, i_1, v_2, i_2 we can write

$$e_1 = u_1 - i_1 Z \tag{6.1.6}$$

$$e_2 = \frac{N_2}{N_1} e_1 = \frac{N_2}{N_1} u_1 - \frac{N_2}{N_1} i_1 Z \tag{6.1.7}$$

$$v_2 = e_2 = \frac{N_2}{N_1} e_1 = \frac{N_2}{N_1} u_1 - i_2 \left(\frac{N_2}{N_1} \right)^2 Z \tag{6.1.8}$$

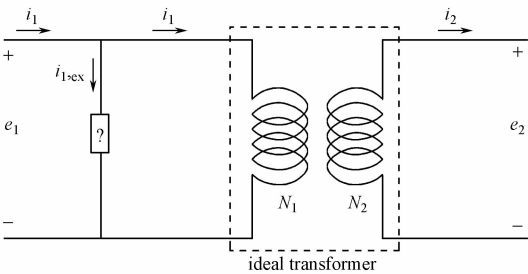
which could describe the circuit in Figure 6.1.4(b). Generally a circuit on a side 1 can be transferred to side 2 by multiplying its component impedances by $(N_2/N_1)^2$, the voltage sources by (N_2/N_1) and the current sources by (N_1/N_2) , while keeping the topology the same.

Equivalent Circuit

To develop the equivalent circuit for a transformer we'll gradually relax the assumptions that we had first imposed. First we'll relax the assumption that the permeability of the iron is infinite. In that case Equation (6.1.4) does not revert to (6.1.5), but rather it becomes

$$N_1 i_1 - N_2 i_2 = R \Phi_m \tag{6.1.9}$$

where R is the reluctance of the path around the core of the transformer and Φ_m the flux



on this path. To preserve the ideal transformer equations as part of our new transformer, we can split i_1 to two components: one i_1' , will satisfy the ideal transformer equation, and the other, $i_{1,ex}$ will just balance the right hand side. Figure 6.1.5 shows this.

$$i_1 = i_1' + i_{1,ex} \tag{6.1.10}$$

$$N_1 i_{1,ex} = R \Phi_m \tag{6.1.11}$$

$$N_1 i_1(t) - N_2 i_2(t) = H \cdot l \tag{6.1.12}$$

We can replace the current source, $i_{1,ex}$, with something simpler if we remember

that the rate of change of flux Φ_m is related to the induced voltage e_1 ^[4]:

$$e_1 = N_1 \frac{d\Phi_m}{dt} = N_1 \frac{d(N_1 i_{1,ex}/R)}{dt} = \left(\frac{N_1^2}{R}\right) \frac{di_{1,ex}}{dt} \tag{6.1.13}$$

Since the current $i_{1,ex}$ flows through something, where the voltage across it is proportional to its derivative, we can consider that this something could be an inductance^[5]. This idea gives rise to the equivalent circuit in Figure 6.1.6, where

$L_m = \frac{N_1^2}{R}$. Let us now relax the assumption that all the flux has to remain in the iron as shown in Figure 6.1.7. Let us call the flux in the iron Φ_m , magnetizing flux, the flux that leaks out of the core and links only coil 1, Φ_{l1} leakage flux 1, and for coil 2, Φ_{l2} leakage flux 2^[6]. Since Φ_{l1} links only coil 1, then it should be related only to the current there, and the same should be true for the second leakage flux.

$$\begin{aligned}\Phi_{l1} &= N_1 i_1 / R_{l1} \\ \Phi_{l2} &= N_2 i_2 / R_{l2}\end{aligned}$$

where R_{l1} and R_{l2} correspond to paths that are partially in the iron and partially in the air. As these currents change, so do the leakage fluxes, and a voltage is induced in each coil:

$$e_1 = \frac{d\lambda_1}{dt} = N_1 \left(\frac{d\Phi_m}{dt}\right) + N_1 \frac{d\Phi_{l1}}{dt} = e_1 + \left(\frac{N_1^2}{R_{l1}}\right) \frac{di_1}{dt} \tag{6.1.14}$$

$$e_2 = \frac{d\lambda_2}{dt} = N_2 \left(\frac{d\Phi_m}{dt}\right) + N_2 \frac{d\Phi_{l2}}{dt} = e_2 + \left(\frac{N_1^2}{R_{l2}}\right) \frac{di_2}{dt} \tag{6.1.15}$$

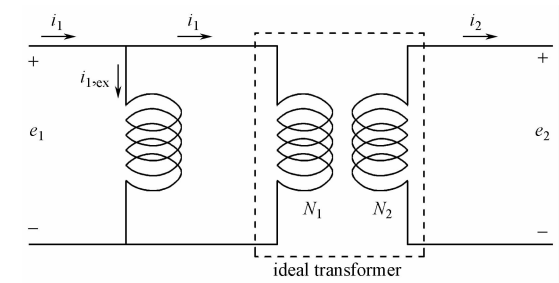


Figure 6.1.6 Ideal transformer plus magnetizing branch.

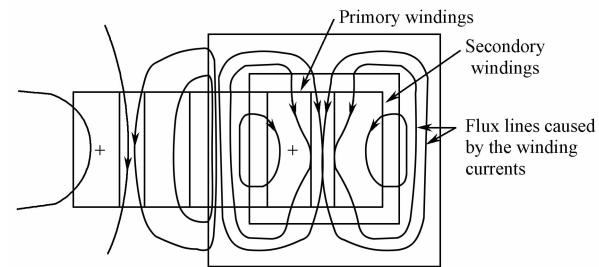


Figure 6.1.7 If the currents in the two windings were to have cancelling values of $N \cdot i$, then the only flux left would be the leakage fluxes. This is the case shown here, designed to point out these fluxes.

If we define, $L_{l1} = \frac{N_1^2}{R_{l1}}, L_{l2} = \frac{N_2^2}{R_{l2}}$, then we can arrive to the equivalent circuit in

Figure 6. 1. 8. To this circuit we have to add:

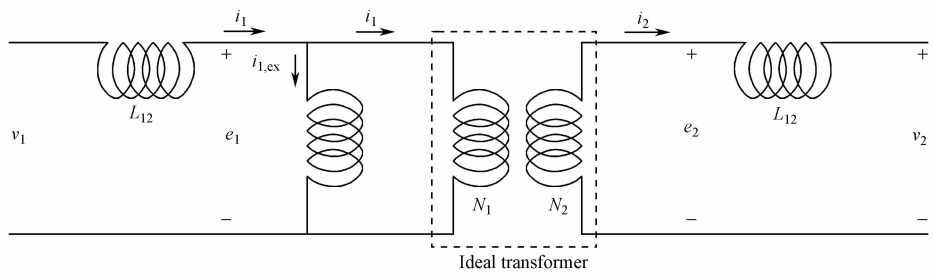


Figure 6. 1. 8 Equivalent circuit of a transformer plus magnetizing and leakage inductances.

1. The winding (ohmic) resistance in each coil, $R_{1,\text{wdg}}$, $R_{2,\text{wdg}}$, with losses, $P_{1,\text{wdg}} = i_1^2 R_{1,\text{wdg}}$, $P_{2,\text{wdg}} = i_2^2 R_{2,\text{wdg}}$.
2. some resistance to represent iron losses. These losses (at least the eddy-current ones) are proportional to the square of the flux. But the flux is proportional to the square of the induced voltage e_1 , hence $P_{\text{iron}} = k e_1^2$. Since this resembles the losses of a resistance supplied by voltage e_1 , we can develop the equivalent circuit Figure 6. 1. 9.

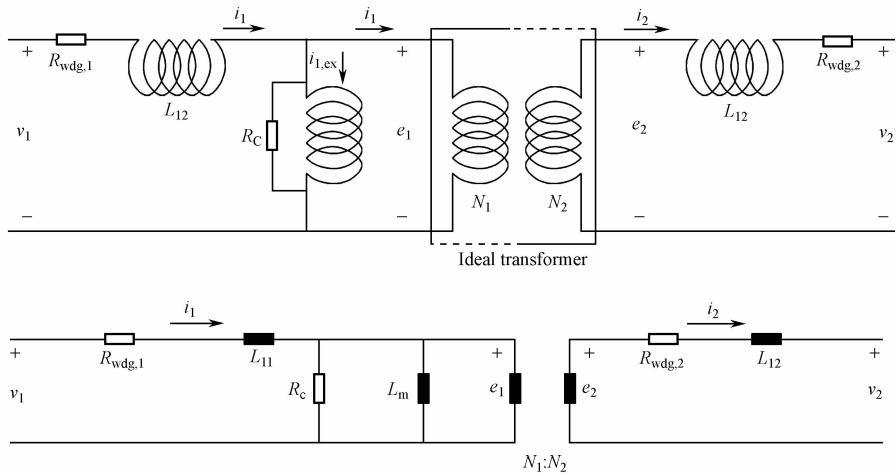


Figure 6. 1. 9 Equivalent circuit for a real transformer.

Losses and Ratings

Again for a given frequency, the power losses in the core (iron losses) increase with the voltage e_1 (or e_2). These losses cannot be allowed to exceed a limit, beyond which the temperature of the hottest spot in the transformer will rise above the point that will decrease dramatically the life of the insulation^[7]. Limits therefore are put to E_1 and E_2 (with a ratio of N_1/N_2), and these limits are the voltage limits of the transformer.

Similarly, winding Joule losses have to be limited, resulting in limits to the currents I_1 and I_2 .

Typically a transformer is described by its rated voltages, E_{1N} and E_{2N} , that give both the limits and turns ratio. The ratio of the rated currents, I_{1N}/I_{2N} , is the inverse of the ratio of the voltages if we neglect the magnetizing current. Instead of the transformer rated currents, a transformer is described by its rated apparent power:

$$S_N = E_{1N}I_{1N} = E_{2N}I_{2N} \quad (6.1.16)$$

Under rated conditions, i. e. maximum current and voltage, in typical transformers the magnetizing current $I_{1,ex}$, does not exceed 1% of the current in the transformer. Its effect therefore on the voltage drop on the leakage inductance and winding resistance is negligible.

Under maximum (rated) current, total voltage drops on the winding resistances and leakage inductances do not exceed in typical transformers 6% of the rated voltage. The effect therefore of the winding current on the voltages E_1 and E_2 is small, and their effect on the magnetizing current can be neglected.

These considerations allow us to modify the equivalent circuit in Fig 6.1.9, to obtain the slightly inaccurate but much more useful equivalent circuits in Figs 6.1.10 (a), (b), and (c).

6.1.2 Specialized English Words

electromechanical	机电的, 电机的	permeability	导磁率
flow	流	flux	磁通
utility pole	电线杆	corollary	必然的结果, 系, 结论
cylinder	圆柱体	flux linkages	磁链
bushings	套管, 衬套	field intensity	场强
core	铁心	reluctance	磁阻
coil	线圈	primary windings	原边绕组
assembly	部件	secondary windings	次边绕组
insulating oil	绝缘油	leakage flux	漏磁通
cutout	截面	magnetizing flux	磁化磁通
tank	箱体	eddy-current	涡流
copper	铜	square	平方
HV bushing	高压套管	joule losses	焦耳损
LV bushing	低压套管	iron losses	铁损
surge suppressor	电涌抑制器	turns	匝
arrester	避雷器		

6.1.3 Notes

[1] They make it possible to increase or decrease the voltage so that power can be transmitted at a voltage level that results in low costs, and can be distributed and used safely. 句中“so that”后面跟的是两个并列的从句“power can be...”, “and can be...”. “that results low costs”是定语从句, 修饰“voltage level”; “make it possible”意为“使……成为可能”。全句译为“变压器能提高或者降低电压, 以实现电能的低成本传输及安全的分配和使用。”

[2] The system of insulation is also associated with that of cooling the core/coil assembly. 句中的“that”指代的是“The system of insulation”, 表示铁心或者线圈的绝缘。全句译为“绝缘系统也通常与铁心或线圈的冷却部件相关。”

[3] We assign dots at one terminal of each coil in the following fashion: if the flux in the core changes, inducing a voltage in the coils, and the dotted terminal of one coil is positive with respect its other terminal, so is the dotted terminal of the other coil. 句中if后面引导的条件状语从句是两个并联句: “if the flux ..., and ...”, 该状语从句中的“inducing”为现在分词状语, 表示结果。“so is the dotted terminal of the other coil”是“so”引导的倒装句。全句译为“我们按如下方式在每个线圈的一端点上标记: 当铁心内的磁通发生变化, 在线圈中产生感应电压时, 线圈带点的一端相对与其另外一端为正。另一线圈也是如此。”

[4] We can replace the current source, $i_{1,ex}$, with something simpler if we remember that the rate of change of flux Φ_m is related to the induced voltage e_1 . “if we ...”为条件状语从句后置, “that the rate of...”是宾语从句, “is related to”这里的意思是“与...有关系”。全句译为“如果我们记得磁通 Φ_m 的变化率是与感应电感 e_1 有关系的, 那么可以用更简单表达式来表示电流源 $i_{1,ex}$ 。”

[5] Since the current $i_{1,ex}$ flows through something, where the voltage across it is proportional to its derivative, we can consider that this something could be an inductance. “since...”为原因状语从句置于句首, “where...its derivative”是同位语从句, 进一步说明“something”, 其中的“where the voltage across it”是名词性主语从句, “it”指代“something”, 后面的“its”也是指代“something”。全句译为“由于电流 $i_{1,ex}$ 流过的元件的两端的电压正比于其微分, 所以我们可以判断其为电感。”

[6] Let us call the flux in the iron Φ_m , magnetizing flux, the flux that leaks out of the core and links only coil 1, Φ_{l1} leakage flux 1, and for coil 2, Φ_{l2} leakage flux 2. “Let us call”后面接了三个并联的谓语“the flux in the iron...”, “the flux that leaks out...”和“for coil...”。全句译为“我们称铁心中的磁通为 Φ_m 为磁化磁通, 称漏出铁心且线圈 1 交链的磁通 Φ_{l1} 为漏磁通 1, 称漏出铁心且线圈 2 交链的磁通 Φ_{l2} 为漏磁通 2。”

[7] These losses cannot be allowed to exceed a limit, beyond which the temperature of the hottest spot in the transformer will rise above the point that will decrease

dramatically the life of the insulation. 句中“beyond which...will rise above the point”为定语从句修饰“limit”; “that the point that will decrease dramatically the life of insulation”又是“the point”的定语从句。全句译为“这种损耗必须限定在一定的范围内, 若超过此范围变压器最热的区域的温度将会超过临界点, 以致绝缘的寿命将会大大缩短。”

6.1.4 Reference Translation

变 压 器

介绍

尽管变压器没有旋转的部件,但是它本质上还是属于机电能量变换设备。变压器能提高或者降低电压,以实现电能的低成本传输及安全的分配和使用。另外,它还能进行匹配阻抗和调节网络中的电力潮流(有功或者无功)。

在电线杆上的变压器我们只能看见一个圆柱体和伸出来的一些电缆。这些电缆通过绝缘套管进入变压器,绝缘套管起隔离电缆与变压器箱体的作用。在箱体内有铁心,上面通常绕有铜制且绝缘的线圈。绝缘系统也常与铁心或线圈的冷却部件相关。通常采用纸来绝缘,整个部件都浸泡在绝缘油中,绝缘油可以在增强纸的绝缘强度的同时,将铁心-绕组部件的热量由箱体传递到空气中。图 6.1.1 画出了典型的配电变压器的截面图。

理想变压器

人类制造的东西几乎没有理想的,变压器也不例外。理想变压器基于非常简单的概念和大量的假设。这是我们在高中学到的变压器。

如图 6.1.2 所示,假设铁心线圈导磁率为无限大,两个匝数分别为 N_1 和 N_2 的线圈(电阻为零)绕在铁心上面。所有的磁通都经过铁心。我们按如下方式在每个线圈的一端点上标记:当铁心内的磁通发生变化,在线圈中产生感应电压时,线圈带点的一端相对于其另外一端为正。另一线圈也是如此。进而言之,流入带点端的电流产生的磁通方向相同。

假设铁心中建立了某个随时间变化的磁通 $\Phi(t)$, 那么在每个线圈中的磁通链为 $\lambda_1 = N_1\Phi(t)$ 和 $\lambda_2 = N_2\Phi(t)$ 。在两个线圈中感应的电压为

$$e_1(t) = \frac{d\lambda_1}{dt} = N_1 \frac{d\Phi}{dt} \tag{6.1.1}$$

$$e_2(t) = \frac{d\lambda_2}{dt} = N_2 \frac{d\Phi}{dt} \tag{6.1.2}$$

两式相除可得

$$\frac{e_1(t)}{e_2(t)} = \frac{N_1}{N_2} \tag{6.1.3}$$

另一方面,流过线圈的电流和电场强度 H 有一定的关系。如果电流 i_1 按照所指的方向流入线圈 1 的带点的一端,电流 i_2 从线圈 2 带点的一端流出,那么有

$$N_1 \cdot i_1(t) - N_2 \cdot i_2(t) = H \cdot l \tag{6.1.4}$$

但是 $B = \mu_{\text{iron}} H$, 且 B 是有限的而 μ_{iron} 是无限的,因此 $H=0$ 。我们知道实际中这是不可

能的,所以只存在于理想变压器。

最后有

$$\frac{i_1}{i_2} = \frac{N_2}{N_1} \quad (6.1.5)$$

式(6.1.3)和式(6.1.5)描述的是理想变压器,它是一个二端口网络。由这两个公式所定义的网络符号见图 6.1.3。理想变压器有一个有趣的特性。下面我们将讨论用等值电路替代一个包含变压器的二端口网络。图 6.1.4(a)所示电路为一个变量为 v_1, i_1, v_2, i_2 的二端口网络,可用以下公式来描述图 6.1.4(b)所示电路:

$$e_1 = u_1 - i_1 Z \quad (6.1.6)$$

$$e_2 = \frac{N_2}{N_1} e_1 = \frac{N_2}{N_1} u_1 - \frac{N_2}{N_1} i_1 Z \quad (6.1.7)$$

$$v_2 = e_2 = \frac{N_2}{N_1} e_1 = \frac{N_2}{N_1} u_1 - i_2 \left(\frac{N_2}{N_1} \right)^2 Z \quad (6.1.8)$$

通常在保持电路拓扑图不变的情况下,将原边阻抗乘以 $(N_2/N_1)^2$,电压源乘以 N_2/N_1 ,电流源乘以 N_1/N_2 ,可以将原边传送到副边。

等值电路

在推导变压器的等效电路时,我们将逐步取消我们开始所做的一些假设。首先我们不再认为铁心的导磁率无穷大。因此式(6.1.4)不能转换为式(6.1.5),而应变成

$$N_1 i_1 - N_2 i_2 = R \Phi_m \quad (6.1.9)$$

式中 R 是铁心通路的磁阻, Φ_m 是通路中的磁通。为了在新变压器中保留理想变压器的公式,我们将 i_1 分成两部分:一部分是 i'_1 ,适用于理想变压器的公式,另一部分是 $i_{1,\text{ex}}$,用于平衡右边电路。如图 6.1.5 所示。

$$i_1 = i'_1 + i_{1,\text{ex}} \quad (6.1.10)$$

$$N_1 i_{1,\text{ex}} = R \Phi_m \quad (6.1.11)$$

$$N_1 i_1(t) - N_2 i_2(t) = H \cdot l \quad (6.1.12)$$

如果我们记得磁通 Φ_m 的变化率是与感应电感 e_1 有关系的,那么可以用更简单的表达式来表示电流源 $i_{1,\text{ex}}$:

$$e_1 = N_1 \frac{d\Phi_m}{dt} = N_1 \frac{d(N_1 i_{1,\text{ex}}/R)}{dt} = \left(\frac{N_1^2}{R} \frac{di_{1,\text{ex}}}{dt} \right) \quad (6.1.13)$$

由于电流 $i_{1,\text{ex}}$ 流过的元件的两端的电压正比于其微分,所以我们可以判断它为电感。这样可以得到图 6.1.6 的等效电路,其中 $L_m = \frac{N_1^2}{R}$ 。现在我们取消磁通全部在铁心中的假设,如图 6.1.7 所示。我们称铁心中的磁通为 Φ_m 为磁化磁通,称漏出铁心且线圈 1 交链的磁通 Φ_{l1} 为漏磁通 1,称漏出铁心且线圈 2 交链的磁通 Φ_{l2} 为漏磁通 2。由于 Φ_{l1} 仅与线圈 1 交链,所以它应仅与该位置的电流有关,漏磁通 2 同样如此。

$$\Phi_{l1} = N_1 i_1 / R_{l1}$$

$$\Phi_{l2} = N_2 i_2 / R_{l2}$$

式中 R_{l1} 和 R_{l2} 对应于部分在铁心部分在空气中的通路。当电流变化时,漏磁通也发生变化,在每个线圈中也产生感应电压:

$$e_1 = \frac{d\lambda_1}{dt} = N_1 \left(\frac{d\Phi_m}{dt} \right) + N_1 \frac{d\Phi_{l1}}{dt} = e_1 + \left(\frac{N_1^2}{R_{l1}} \right) \frac{di_1}{dt} \quad (6.1.14)$$

$$e_2 = \frac{d\lambda_2}{dt} = N_2 \left(\frac{d\Phi_m}{dt} \right) + N_2 \frac{d\Phi_{l2}}{dt} = e_2 + \left(\frac{N_2^2}{R_{l2}} \right) \frac{di_2}{dt} \quad (6.1.15)$$

若定义 $L_{l1} = \frac{N_1^2}{R_{l1}}, L_{l2} = \frac{N_2^2}{R_{l2}}$, 我们可以得到如图 6.1.8 所示的等效电路。对于该电路我们还必须加两点:

1. 每个线圈的绕线电阻(欧姆), $R_{1,wdg}, R_{2,wdg}$, 其损耗为 $P_{1,wdg} = i_1^2 R_{1,wdg}, P_{2,wdg} = i_2^2 R_{2,wdg}$ 。
2. 用来表示铁损的电阻。这些损耗(至少包括涡流损耗)正比于磁通平方。而磁通正比于感应电压 e_1 的平方,因此 $P_{iron} = k e_1^2$ 。由于这些类似的电阻损耗是由电压 e_1 产生的,因此我们可以得到如图 6.1.9 所示的等值电路。

损耗和额定值

同样,对于给定的频率,在线圈中的功率损耗(铁损)随着电压 e_1 (或 e_2) 的增加而增加。这种损耗必须限定在一定的范围内,若超过此范围变压器最热的区域的温度将会超过临界点,以至于绝缘的寿命将会大大缩短。因此将这个范围用到 E_1 和 E_2 (按照 N_1/N_2 的比例),即为变压器的电压临界值。

类似地,绕组的焦耳损也有临界值,这将体现在电流 I_1 和 I_2 的临界值上。

通常我们用额定电压 E_{1N} 和 E_{2N} 来描述变压器,它表示了临界值和匝数比。若我们忽略磁化电流,则额定电流的比值 I_{1N}/I_{2N} 为电压比值的倒数。除了额定电流,变压器也可以用它的额定视在功率来描述:

$$S_N = E_{1N} I_{1N} = E_{2N} I_{2N} \quad (6.1.16)$$

在额定值的前提下,也就是说,在最大电流和电压情况下,变压器的典型磁化电流 $I_{1,ex}$ 不会超过变压器电流的 1%。因此它对于由漏感产生的电压降和绕组电阻方面的影响可以忽略。在典型的变压器中,在最大的额定电流情况下,由绕组电阻和漏感压降产生的全部压降不会超过额定电压的 6%。因此绕组电流对电压 E_1 和 E_2 的影响很小,且对磁化电流的影响可以忽略。

根据上述考虑,我们将图 6.1.9 的等效电路图修改为图 6.1.10(a)、(b)和(c),它们的准确性略微下降,但更加有用。

6.1.5 Reading Materials

Transformer Tests

We are usually given a transformer, with its frequency, power and voltage ratings, but without the values of its impedances. It is often important to know these

impedances, in order to calculate voltage regulation, efficiency etc. , in order to evaluate the transformer (e. g. if we have to choose from many) or to design a system. Here we'll work on finding the equivalent circuit of a transformer, through two tests. We'll use the results of these test in the per-unit system.

First we notice that if the relative values are as described in section 6. 1. 1. 4, we cannot separate the values of the primary and secondary resistances and reactances. We will lump $R_{1,\text{wdg}}$ and $R_{2,\text{wdg}}$ together, as well as X_{l1} and X_{l2} . This will leave four quantities to be determined, R_{wdg} , X_l , R_c and X_m .

Open Circuit Test

We leave one side of the transformer open circuited, while to the other we apply rated voltage (i. e. $V_{\text{oc}} = 1 \text{ pu}$) and measure current and power. On the open circuited side of the transformer rated voltage appears, but we just have to be careful not to close the circuit ourselves. The current that flows is primarily determined by the impedances X_m and R_c , and it is much lower than rated. It is reasonable to apply this voltage to the low voltage side, since (with the ratings of the transformer in our example) is it easier to apply 120 V, rather than 4000 V. We will use these two measurements to calculate the values of R_c and X_m .

Dropping the subscript *pu*, using the equivalent circuit of Fig 6. 1. 11 (b) and neglecting the voltage drop on the horizontal part of the circuit, we calculate

$$P_{\text{oc}} = \frac{V_{\text{oc}}^2}{R_c} = \frac{1}{RC} \quad (6.1.17)$$

$$I_{\text{oc}} = \frac{V_{\text{oc}}}{R_c} + \frac{V_{\text{oc}}}{jX_m}$$

$$I_{\text{oc}} = 1 \sqrt{\frac{1}{R_c^2} + \frac{1}{X_m^2}} \quad (6.1.18)$$

Equations (6.1.17) and (6.1.18), allow us to use the results of the short circuit test to calculate the vertical (core) branch of the transformer equivalent circuit.

Short Circuit Test

To calculate the remaining part of the equivalent circuit, i. e. the values of R_{wdg} and X_l , we short circuit one side of the transformer and apply rated current to the other. We measure the voltage of that side and the power drawn. On the other side, (the short-circuited one) the voltage is of course zero, but the current is rated. We often apply voltage to the high voltage side, since (a) the applied voltage need not be high and (b) the rated current on this side is low.

Using the equivalent circuit of Fig 6. 1. 11(a), we notice that

$$P_{\text{sc}} = I_{\text{sc}}^2 R_{\text{wdg}} = 1 \cdot R_{\text{wdg}} \quad (6.1.19)$$

$$V_{\text{sc}} = I_{\text{sc}} (R_{\text{wdg}} + jX_l)$$

$$V_{sc} = 1 \cdot \sqrt{R_{wdg}^2 + X_l^2} \quad (6.1.20)$$

Equations (6.1.19) and (6.1.20) can give us the values of the parameters in the horizontal part of the equivalent circuit of a transformer.

6.2 DC Motors and AC Motors

6.2.1 Text

DC Machines

DC machines have faded from use due to their relatively high cost and increased maintenance requirements. Nevertheless, they remain good examples for electromechanical systems used for control. We'll study DC machines here, at a conceptual level, for two reasons:

1. DC machines although complex in construction, can be useful in establishing the concepts of emf and torque development, and are described by simple equations.
2. The magnetic fields in them, along with the voltage and torque equations can be used easily to develop the ideas of field orientation.

In doing so we will develop basic steady-state equations, again starting from fundamentals of the electromagnetic field. We are going to see the same equations in 'Brushless DC' motors, when we discuss synchronous AC machines.

Geometry, Fields, Voltages, and Currents

Let us start with the geometry shown in Figure 6.2.1.

This geometry describes an outer iron window (stator), through which (i. e. its center part) a uniform magnetic flux is established, say $\hat{\Phi}$. How this is done (a current in a coil, or a permanent magnet) is not important here.

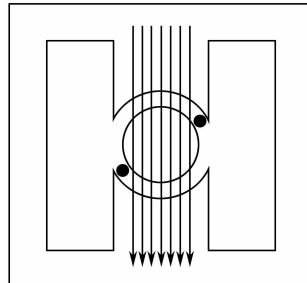


Figure 6.2.1 Geometry of an elementary DC motor.

In the center part of the window there is an iron cylinder (called rotor), free to rotate around its axis. A coil of one turn is wound diametrically around the cylinder, parallel to its axis^[1]. As the cylinder and its coil rotate, the flux through the coil changes. Figure 6.2.2 shows consecutive locations of the rotor and we can see that the flux through the coil changes both in value and direction. The top graph of Figure 6.2.3 shows how the flux linkages of the coil through the coil would change, if the rotor were to rotate at a constant angular velocity, ω ^[2].

$$\lambda = \hat{\Phi} \cos[\omega t] \quad (6.2.1)$$

Since the flux linking the coil changes with time, then a voltage will be induced in this coil, v_{coil} ,

$$v_{\text{coil}} = \frac{d\lambda}{dt} = -\hat{\Phi} \omega \sin(\omega t) \quad (6.2.2)$$

shown in the second graph of Figure 6.2.3. The points marked there correspond to the position of the rotor in Figure 6.2.2.

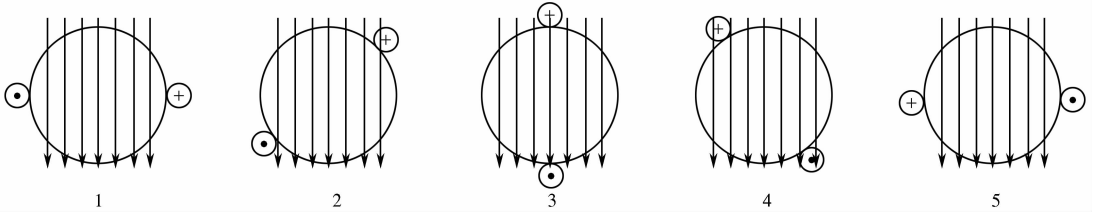


Figure 6.2.2 Flux through a coil of a rotating DC machine.

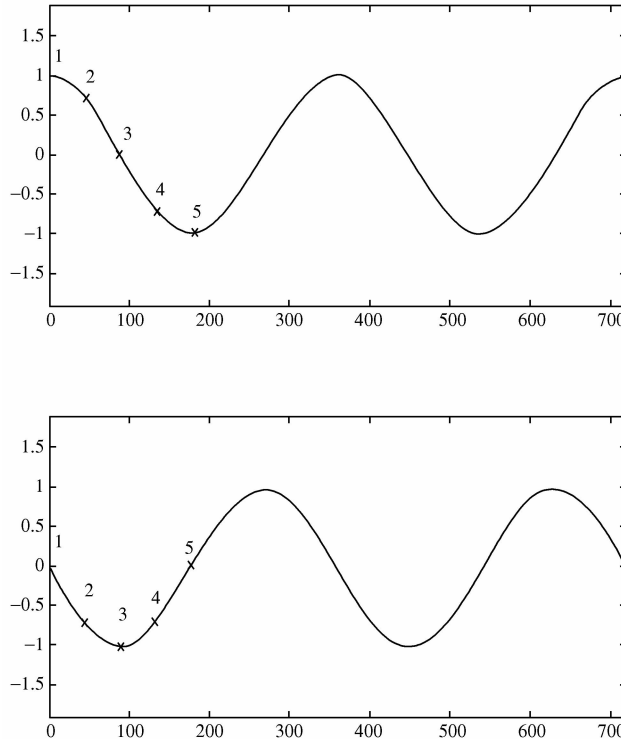


Figure 6.2.3 Flux and voltage in a coil of the DC machine in Figure 6.2.2. Points 1-5 represent the coil positions.

This alternating voltage has to somehow be rectified, since this is a DC machine. Although this can be done electronically, a very old mechanical method exists. The coil is connected not to the DC source or load, but to two ring segments, solidly attached to it and the rotor, and hence rotating with it^[3]. Two ‘brushes’, i. e. conducting pieces of material (often carbon/copper) are stationary and sliding on these ring segments as shown in Figure 6. 2. 4.

The structure of the ring segments is called a commutator. As it rotates, the brushes make contact with the opposite segments just as the induced voltage goes through zero and switches sign.

Figure 6. 2. 5 shows the induced voltage and the terminal voltage seen at the voltmeter of Figure 6. 2. 4. If a number of coils are placed on the rotor, as shown in Figure 6. 2. 6, each connected to a commutator segment, the total induced voltage to the coils, E will be

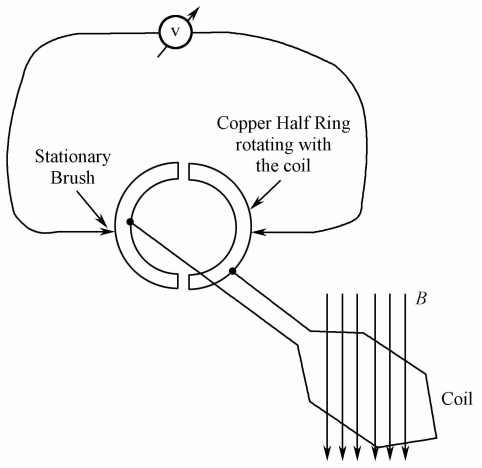


Figure 6. 2. 4 A coil of a DC motor and a commutator with brushes.

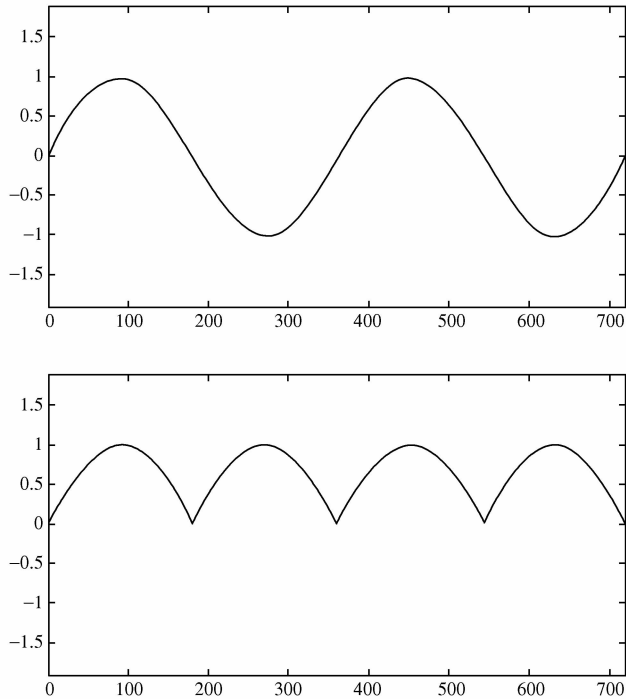


Figure 6. 2. 5 Induced voltage in a coil and terminal voltage in an elementary DC machine.

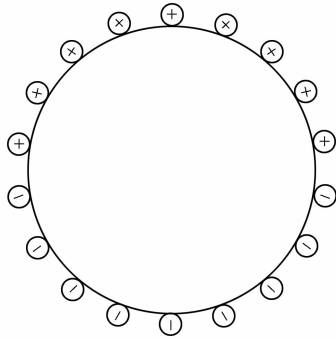


Figure 6. 2. 6 Coils on the rotor of DC machine.

$$E = k \hat{\Phi} \omega \quad (6. 2. 3)$$

where k is proportional to the number of coils.

Due to,

$$E \cdot i = T \omega \quad (6. 2. 4)$$

$$k \hat{\Phi} \omega i = T \omega \quad (6. 2. 5)$$

$$T = k \hat{\Phi} i \quad (6. 2. 6)$$

If the electrical machine is connected to a load or a source as in Figure 6. 2. 7, the induced voltage and terminal voltage

will be related by

$$V_{\text{term}} = E - i_g R_{\text{wdg}} \quad \text{for a generator} \quad (6. 2. 7)$$

$$V_{\text{term}} = E + i_m R_{\text{wdg}} \quad \text{for a motor} \quad (6. 2. 8)$$

Notes

- The field of the DC motor can be created either by a DC current or a permanent magnet.
- These two fields, the one coming from the stator and the one coming from the moving rotor, are both stationary (despite rotation) and perpendicular to each other.
- If the direction of current in the stator and in the rotor reverse together, torque will remain in the same direction. Hence if the same current flows in both windings, it could be AC and the motor will not reverse (e. g. hairdryers, power drills).

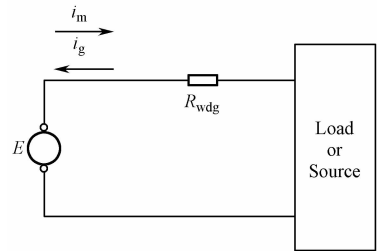


Figure 6. 2. 7 Circuit with a DC machine.

Induction Machines

Induction machines are often described as the ‘workhorse of industry’. This cliché reflects the reality of the qualities of these machines. They are cheap to manufacture, rugged and reliable and find their way in most possible applications^[4]. Variable speed drives require inexpensive power electronics and computer hardware, and allowed induction machines to become more versatile. In particular, vector or field-oriented control allows induction motors to replace DC motors in many applications.

The stator of an induction machine is a typical three-phase one, as described in the previous chapter. The rotor can be one of two major types. Either (a) it is wound in a

fashion similar to that of the stator with the terminals led to slip rings on the shaft, as shown in Figure 6.2.8, or (b) it is made with shorted bars. Figure 6.2.9 shows the rotor of such a machine, while Figure 6.2.10 show the shorted bars and the laminations.

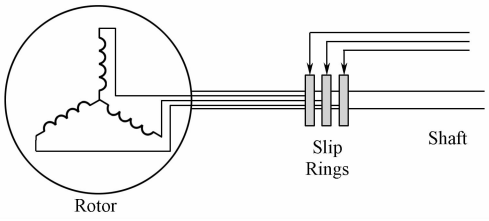


Figure 6.2.8 Wound rotor slip rings and connections.

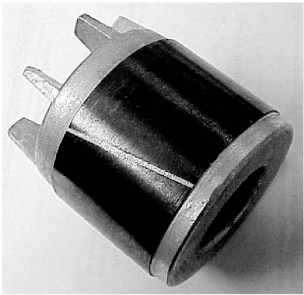


Figure 6.2.9 Rotor for squirrel cage induction motor.

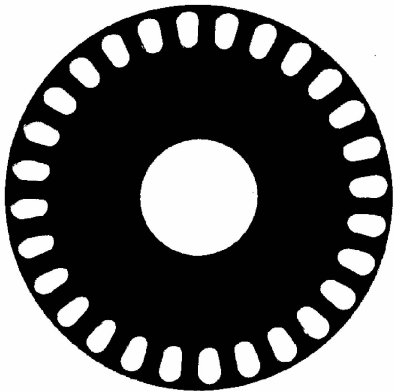
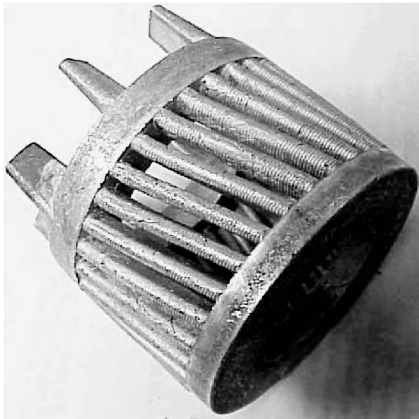


Figure 6.2.10 Rotor components of a squirrel cage induction motor.

The picture of the rotor bars is not easy to obtain, since the bars are formed by casting aluminum in the openings of the rotor laminations. In this case the iron laminations were chemically removed.

As these rotor windings or bars rotate within the magnetic field created by the stator magnetizing currents, voltages are induced in them. Figure 6.2.11 shows the rotor bars and their voltages. If the rotor were to stand still, then the induced voltages would be very similar to those induced in the stator windings^[5]. In the case of squirrel cage rotor, the voltage induced in the bars will be slightly out of phase with the voltage in the next one, since the flux linkages will change in it after a short delay^[6].

If the rotor is moving at synchronous speed, together with the field, no voltage will be induced in the bars or the windings.

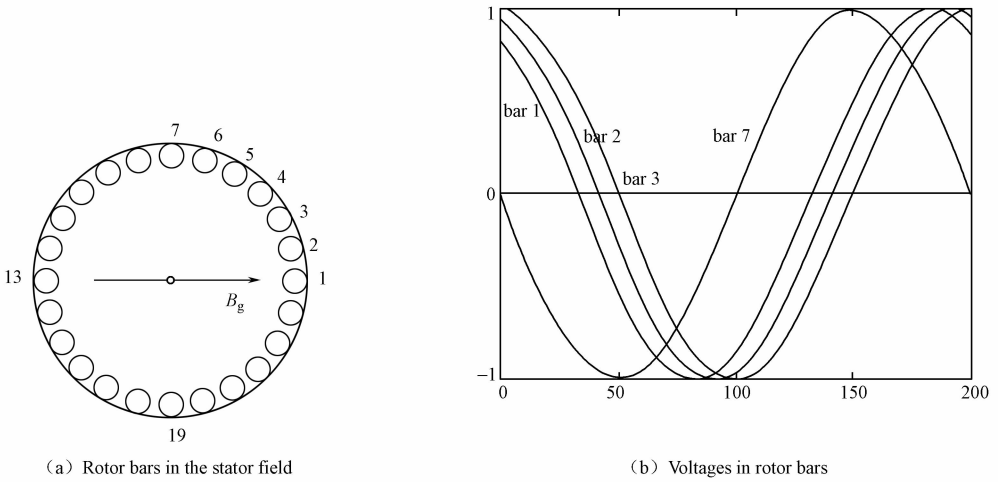


Figure 6.2.11 Rotor bars and their voltages.

Generally when the synchronous speed is $\omega_s = 2\pi f_s$, and the rotor speed ω_0 , the frequency of the induced voltages will be f_r , where $2\pi f_r = \omega_s - \omega_0$. Maxwell's equation becomes here:

$$\epsilon = v \times B_g \quad (6.2.9)$$

where v is the relative velocity of the rotor with respect to the field,

$$v = (\omega_s - \omega_0)r \quad (6.2.10)$$

Since a voltage is induced in the bars, and these are short-circuited, currents will flow in them. The current density $J(\theta)$ will be

$$J(\theta) = \frac{1}{\rho} \epsilon \quad (6.2.11)$$

These currents are out of phase in different bars, just like the induced voltages. To simplify the analysis we can consider the rotor as one winding carrying currents sinusoidally distributed in space. This will be clearly the case for a wound rotor. It will also be the case for uniformly distributed rotor bars, but now each bar, located at an angle θ will carry different current, as shown in Figure 6.2.12(a),

$$J = \frac{1}{\rho} (\omega_s - \omega_0) \cdot B_g(\theta) \quad (6.2.12)$$

$$J(\theta) = \frac{1}{\rho} (\omega_s - \omega_0) \cdot \hat{B}_g \sin(\theta) \quad (6.2.13)$$

We can replace the bars with a conductive cylinder as shown in Figure 6.2.12(b). We define as slip s the ratio:

$$s = \frac{\omega_s - \omega_0}{\omega_s} \quad (6.2.14)$$

At starting the speed is zero, hence $s = 1$, and at synchronous speed, $\omega_s = \omega_0$, hence $s = 0$. Above synchronous speed $s < 0$, and when the rotor rotates in a direction opposite of the magnetic field $s > 1$.

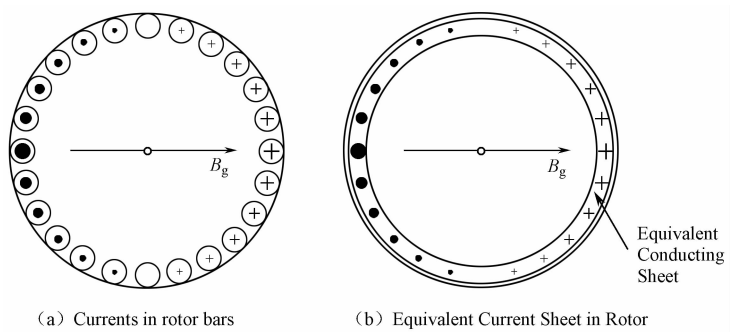


Figure 6.2.12 Current Distribution in equivalent conducting sheet.

We already determined that the voltages induced in the rotor bars are of slip frequency, $f_r = (\omega_s - \omega_0)/2\pi$. At rotor speeds near synchronous, this frequency, f_r is quite small. The rotor bars in a squirrel cage machine possess resistance and leakage inductance, but at very low frequencies, i. e. near synchronous speed, we can neglect this inductance. The rotor currents therefore are limited near synchronous speed by the rotor resistance only.

6.2.2 Specialized English Words

motor	电动机,电机,马达	flux linkage	磁链,磁通匝连数
electromechanical	机电的	angular velocity	角速度
emf	电动势(electromotive force 之缩写)	workhorse	驮马,重负荷机器,主力,骨干
torque	力矩,转矩	carbon	碳
magnetic fields	磁场	stationary	固定的,不动的
brushless DC	无刷直流电动机	commutator	换向器,整流子
synchronous AC machines	同步交流电机	induced voltage	感应电压
geometry	结构,几何形状	voltmeter	电压表
stator	定子	generator	发电机
cylinder	圆柱体,圆筒,圆柱	manufacture	制造,加工
permanent magnet	永磁体	rugged	耐用的,粗糙的,强壮的
diametrically	作为直径地	field-oriented	场导向的
axis	轴,坐标轴	slip rings	集电环,滑动环,汇流环

laminations 铁心片,叠片结构,迭片
aluminum 铝
Maxwell's equation 麦克斯韦方程
current density 电流密度

slip frequency 转差频率
squirrel cage machine 鼠笼电机
in phase 同相,同相地
out of phase 异相,不同相位

6.2.3 Notes

[1] A coil of one turn is wound diametrically around the cylinder, parallel to its axis. 句中“parallel to its axis”为并列的表语,其中的“its”指代“a cylinder”。全句译为“一个单匝的线圈绕在圆柱体上,且和圆柱体的轴线平行。”

[2] The top graph of Figure 6.2.3 shows how the flux linkages of the coil through the coil would change, if the rotor were to rotate at a constant angular velocity, ω . 句中“how the flux linkages...would change”为宾语从句,“the flux linkages”为宾语从句的主语,“if the rotate...”为非真实的条件状语从句,采用虚拟语气。全句译为“图 6.2.3 中上部的图给出了在假设转子以恒定角速度 ω 旋转的情况下,通过线圈磁链的变化过程。”

[3] The coil is connected not to the DC source or load, but to two ring segments, solidly attached to it and the rotor, and hence rotating with it. 句中“attached to it and the rotor”为过去分词短语做状语,“it”指“two ring segments”;“rotating with it”为现在分词短语,亦做状语。“it”同样指“two ring segments”。全句可译为“线圈不是直接连接到直流电源或者负载上面,而是连接到两段环形片上,和环形片及转子牢固相连,因此线圈和环形片一起旋转。”

[4] They are cheap to manufacture, rugged and reliable and find their way in most possible applications. 句中“rugged”和“reliable”是并列的表语,意思是 They are rugged, they are reliable。“find”则是和“are”并列的谓语动词。全句译为“它们造价低,耐用、可靠,且能在大多数应用场合使用。”

[5] If the rotor were to stand still, then the induced voltages would be very similar to those induced in the stator windings. 该句为“if”引起的与现在事实相反的虚拟语气。“induced in the stator windings”为过去分词短语修饰“those”,“those”指代“voltages”。全句译为“若转子静止不动,则转子内产生的感应电压将和定子产生的感应电压非常类似。”

[6] In the case of squirrel cage rotor, the voltage induced in the bars will be slightly out of phase with the voltage in the next one, since the flux linkages will change in it after a short delay. 句中“induced in bars”为过去分词做定语说明“voltage”,“since...”为原因状语从句。全句译为“若采用的是鼠笼转子,由于磁链会略有滞后,因此在导条内产生的感应电压将和下一个产生的电压略微有相位差。”

6.2.4 Reference Translation

直流电机与交流电机

直流电机

由于直流电机造价相对较高且后期对维护要求会不断增加,所以应用逐渐变少。尽管如此,对于控制的机电系统来说,它仍然是非常好的例子。在这里我们将从概念的层面上学习直流电机,有以下两点原因:

1. 尽管直流电机在结构上复杂,但是它对于建立电动势和转矩的形成的概念是很有用的,而且可以用简单的方程描述。
2. 直流电机中的磁场以及电压和转矩的方程可以很容易地获得场方向的概念。

与此同时,我们同样从基本的电磁场的概念出发,将推导基本的稳态方程。讨论同步交流电机时,我们将看到和直流无刷电机一样的方程。

结构、场、电压和电流

我们由图 6.2.1 所示的结构开始。

从结构上可以看到外面有一个铁框架(定子),穿过其中心部位建有均匀磁通,用 $\hat{\Phi}$ 表示。磁通是如何产生的(来自线圈里面的电流或永磁铁)在这里并不重要。

框架的中间有一个铁圆柱体(称为转子),绕着它的轴自由旋转。一个单匝的线圈径向绕在圆柱体上,且和圆柱体的轴线平行。当柱体和线圈旋转时,通过线圈的磁通发生改变。图 6.2.2 给出了转子连续的位置图,可以看到通过线圈的磁通在大小和方向上都有改变。图 6.2.3 中上部的图给出了在假设转子以恒定角速度 ω 旋转的情况下,通过线圈磁链的变化过程。

$$\lambda = \hat{\Phi} \cos[\omega t] \quad (6.2.1)$$

由于线圈的磁通链随时间改变,因此在线圈中必定产生感应电压 v_{coil} ,

$$v_{\text{coil}} = \frac{d\lambda}{dt} = -\hat{\Phi}\omega \sin(\omega t) \quad (6.2.2)$$

如图 6.2.3 的第二个图所示。图中的点表示对应于图 6.2.2 中的转子的位置。

由于这是直流电机,因此这个交变的电压必须按照某种方式整流。尽管能用电子的方式来实现,但古老的机械方式仍然沿用。线圈不是直接连接到直流电源或者负载上面的,而是连接到两段环形片上,和环形片及转子牢固相连,因此线圈和环形片一起旋转。由导电材料(通常为碳或铜)构成的电刷是固定不动的,如图 6.2.4 所示滑靠在这些环形片上。

环形片的构造我们称之为换向器。在旋转过程中,电刷与相对的环形片的接触发生在感应电压过零且符号发生改变的时候。

图 6.2.5 给出了由图 6.2.4 中电压表读出的感应电压及端电压。如果转子中安放有大量的线圈,如图 6.2.6 所示,每个线圈都连接到换向片上,那么整个线圈的感应电压 E 将为

$$E = k\hat{\Phi}\omega \quad (6.2.3)$$

式中 k 正比于线圈的数量。

由于

$$E \cdot i = T\omega \quad (6.2.4)$$

$$k\hat{\Phi}\omega i = T\omega \quad (6.2.5)$$

$$T = k\hat{\Phi}i \quad (6.2.6)$$

如图 6.2.7 所示,如果电机连接到一个负载或者电源,那么感应电压和端电压有如下关系:

$$V_{\text{term}} = E - i_{\text{g}} R_{\text{wdg}} \quad \text{发电机} \quad (6.2.7)$$

$$V_{\text{term}} = E + i_{\text{m}} R_{\text{wdg}} \quad \text{电动机} \quad (6.2.8)$$

注:

- 直流电机的磁场可由直流电流或者永磁体建立。
- 有两个磁场,一个是定子产生的,一个是转子产生的,两者都是静止的(尽管有旋转),互相垂直。
- 若定子的电流方向和转子的电流方向一起改变,转矩方向不变。因此如果两个绕组内的电流方向相同,那么该电流应为交流,电机不会反转(如电吹风、电钻)。

感应电机

我们常称感应电机为“工业之老黄牛”。这个老套的比喻反映了这类机器的实际特性。它们易于制造,耐用、可靠,且能在大多数应用场合使用。廉价的电力电子和计算机硬件即可满足调速的要求,这使得感应电机应用更为广泛。尤其是矢量控制和磁场定向控制使得感应电机在许多应用中取代直流电机。

如前面一章所述,感应电机的定子是典型的三相结构。转子有两种主要类型。或者 a) 和定子的绕线方式类似,绕组端子连接到轴上的滑环处,如图 6.2.8 所示;或者 b) 由短路导条组成。图 6.2.9 展示了这种类型的转子的电动机,而图 6.2.10 则给出了短路导条和叠片。

由于转子导条是在敞开的转子叠片上浇铸而成的,因此转子导条的图片很难得到。在这里,是用化学方法将铁心的叠片去掉的。

由于转子的绕组或者导条在由定子磁化电流建立的磁场中旋转,因此会产生感应电压。图 6.2.11 给出了转子导条及它们的电压。若转子静止不动,则转子内产生的感应电压将和定子产生的感应电压非常类似。若采用的是鼠笼转子,由于磁链会略有滞后,因此在导条内产生的感应电压将和下一个产生的电压略微有相位差。

若转子以和磁场同步的速度一起旋转,则在导条或绕组内不会产生电压。

一般来说,当同步速度为 $\omega_s = 2\pi f_s$ 时,转子速度为 ω_0 ,感应电压的频率为 f_r ,这里 $2\pi f_r = \omega_s - \omega_0$ 。麦克斯韦方程则变为

$$\epsilon = v \times B_g \quad (6.2.9)$$

这里 v 是转子相对与磁场的速度,

$$v = (\omega_s - \omega_0)r \quad (6.2.10)$$

由于是在导条内产生的电压,而这些导条是短路的,因此将有电流在里面流动。电流密度 $J(\theta)$ 为

$$J(\theta) = \frac{1}{\rho} \varepsilon \quad (6.2.11)$$

和感应电压一样,不同的导条内的电流是异相的。为简化分析,我们将转子视为一个绕组,绕组内的电流在空间上呈正弦分布。这将清楚地展现出绕线式转子的情况。同样,对于均匀分布的转子导条来说也适用,不过对于每个角度为 θ 的导条来说,所通过的电流是不一样的,如图 6.1.12(a)所示:

$$J = \frac{1}{\rho} (\omega_s - \omega_0) \cdot B_g(\theta) \quad (6.2.12)$$

$$J(\theta) = \frac{1}{\rho} (\omega_s - \omega_0) \cdot \hat{B}_g \sin(\theta) \quad (6.2.13)$$

我们可用如图 6.2.12(b)所示的导电的圆柱体来代替导条。

我们定义转差率 s 为

$$s = \frac{\omega_s - \omega_0}{\omega_s} \quad (6.2.14)$$

起动的时候转速为零,因此 $s=1$, 如果以同步速度旋转, $\omega_s = \omega_0$, 因此 $s=0$ 。大于同步速度则 $s<0$, 当转子按照与磁场相反的方向旋转时,则 $s > 1$ 。

我们已经确定转子导条的感应电压的转差频率为 $f_r = (\omega_s - \omega_0)/2\pi$ 。当转子速度接近同步速度时,这个频率 f_r 是相当小的。在鼠笼电机内转子导条有一定的阻值和漏感,但在频率很低的情况下,即接近同步速度的情况下,我们可以忽略这个电感。因此在接近同步转速时,转子电流仅与转子电阻相关。

6.3 Adjustable Speed Drives

6.3.1 Text

By definition, adjustable speed drives of any type provide a means of variably changing speed to better match operating requirements. Such drives are available in mechanical, fluid and electrical types.

The most common mechanical versions use combinations of belts and sheaves, or chains and sprockets, to adjust speed in set, selectable ratios—2 : 1, 4 : 1, 8 : 1 and so forth. Traction drives, a more sophisticated mechanical control scheme, allow incremental speed adjustments. Here, output speed is varied by changing the contact points between metallic disks, or between balls and cones.

Adjustable speed fluid drives provide smooth, stepless adjustable speed control. There are three major types. Hydrostatic drives use electric motors or internal

combustion engines as prime movers in combination with hydraulic pumps, which in turn drive hydraulic motors. Hydrokinetic and hydroviscous drives directly couple input and output shafts. Hydrokinetic versions adjust speed by varying the amount of fluid in a vortex that serves as the input-to-output coupler. Hydroviscous drives, also called oil shear drives, adjust speed by controlling oil-film thickness, and therefore slippage, between rotating metallic disks.

An eddy current drive, while technically an electrical drive, nevertheless functions much like a hydrokinetic or hydroviscous fluid drive in that it serves as a coupler between a prime mover and driven load^[1]. In an eddy current drive, the coupling consists of a primary magnetic field and secondary fields created by induced eddy currents. The amount of magnetic slippage allowed among the fields controls the driving speed.

In most industrial applications, mechanical, fluid or eddy current drives are paired with constant-speed electric motors. On the other hand, solid state electrical drives (also termed electronic drives), create adjustable speed motors, allowing speeds from zero RPM to beyond the motor's base speed. Controlling the speed of the motor has several benefits, including increased energy efficiency by eliminating energy losses in mechanical speed changing devices. In addition, by reducing, or often eliminating, the need for wear-prone mechanical components, electrical drives foster increased overall system reliability, as well as lower maintenance costs^[2]. For these and other reasons, electrical drives are the fastest growing type of adjustable speed drive.

There are two basic drive types related to the type of motor controlled-DC and AC. A DC direct current drive controls the speed of a DC motor by varying the armature voltage (and sometimes also the field voltage). An alternating current drive controls the speed of an AC motor by varying the frequency and voltage supplied to the motor.

DC Drives

Direct current drives are easy to apply and technologically straightforward. They work by rectifying AC voltage from the power line to DC voltage, then feeding adjustable voltage to a DC motor. With permanent magnet DC motors, only the armature voltage is controlled. The more voltage supplied, the faster the armature turns. With wound-field motors, voltage must be supplied to both the armature and the field. In industry, the following three types of DC drives are most common, as shown in Figure 6.3.1:

Drives: These are named for the silicon controlled rectifiers (also called thyristors) used to convert AC to controlled voltage DC. Inexpensive and easy to use, these drives come in a variety of enclosures, and in unidirectional or reversing styles.



A general-purpose DC SCR drives family.
From left, NEMA 4/12 “totally enclosed”
version, chassis-mount,
NEMA 1 “open” enclosure.

Figure 6.3.1 A general-purpose DC SCR drives family.

Regenerative SCR Drives: Also called four quadrant drives, these allow the DC motor to provide both motoring and braking torque. Power coming back from the motor during braking is regenerated back to the power line and not lost.

Pulse Width Modulated DC Drives: Abbreviated PWM and also called, generically, transistorized DC drives, these provide smoother speed control with higher efficiency and less motor heating. Unlike SCR drives, PWM types have three elements. The first converts AC to DC, the second filters and regulates the fixed DC voltage, and the third controls average voltage by creating a stream of variable width DC pulses. The filtering section and higher level of control modulation account for the PWM drive’s improved performance compared with a common SCR drive^[3].

AC Drives

AC drive operation begins in much the same fashion as a DC drive. Alternating line voltage is first rectified to produce DC. But because an AC motor is used, this DC voltage must be changed back, or inverted, to an adjustable-frequency alternating voltage. The drive’s inverter section accomplishes this. In years past, this was accomplished using SCRs. However, modern AC drives use a series of transistors to invert DC to adjustable-frequency AC. An example is shown in Figure 6.3.2.

This synthesized alternating current is then fed to the AC motor at the frequency and voltage required to produce the desired motor speed. For example, a 60 Hz synthesized frequency, the same as standard line frequency in the United States, produces 100% of rated motor speed. A lower frequency produces a lower speed, and a higher frequency a higher speed. In this way, an AC drive can produce motor speeds from, approximately, 15 to



Figure 6.3.2 With advances in power electronics, even so-called “micro” drives can be used with motors 40 HP or higher. Full-featured unit shown includes keypad programming and alphanumeric display.

200% of a motor’s normally rated RPM — by delivering frequencies of 9 Hz to 120 Hz, respectively.

Today, AC drives are becoming the systems of choice in many industries. Their use of simple and rugged three-phase induction motors means that AC drive systems are the most reliable and least maintenance prone of all. Plus, microprocessor advancements



Figure 6. 3. 3 Encoders can be added to inverter-duty three-phase motors for use in closed-loop vector drive systems.

have enabled the creation of so-called vector drives, which provide greatly enhance response, operation down to zero speed and positioning accuracy^[4]. Vector drives, especially when combined with feedback devices such as tachometers, encoders and resolvers in a closed-loop system, are continuing to replace DC drives in demanding applications^[5]. An Exaple is shown in

Figure 6. 3. 3.

By far the most popular AC drive today is the pulse width modulated type. Though originally developed for smaller-horsepower applications, PWM is now used in drives of hundreds or even thousands of horsepower—as well as remaining the staple technology in the vast majority of small integral and fractional horsepower “micro” and “sub-micro” AC drives^[6], as shown in Figure 6. 3. 4.

Pulse width modulated refers to the inverter’s ability to vary the output voltage to the motor by altering the width and polarity of voltage pulses. The voltage and frequency are synthesized using this stream of voltage pulses. This is accomplished through microprocessor commands to a series of power semiconductors that serve as on-off switches. Today, these switches are usually IGBTs, or isolated gate bipolar transistors. A big advantage to these devices is their fast switching speed resulting in higher pulse or carrier frequency, which minimizes motor noise. ^[7]



Figure 6. 3. 4 “Sub-micro” drives provide a wide array of features in a very small package.

6. 3. 2 Specialized English Words

DC Drives 直流传动

AC drives 交流传动

means 手段,方法

belts and sheaves 三角皮带轮

chains and sprockets 链轮

traction drives 牵引传动

cones 锥体

fluid drives 流体传动

hydrostatic drives 静液压传动

oil-film 油膜

coupler 耦合器

prime mover 原动力

vortex 涡流

armature 电枢

regenerative 回馈,可再生的

hydraulic drives 液体黏性传动

pumps 泵

hydrokinetic 流体动力传动

slippage 滑移

inverter 逆变器

braking torque 制动转矩

staple 主要的,常用的,大宗生产的

carrier frequency 脉冲载波频率

6.3.3 Notes

[1] An eddy current drive, while technically an electrical drive, nevertheless functions much like a hydrokinetic or hydroviscous fluid drive in that it serves as a coupler between a prime mover and driven load. 句中“while technically an electrical drive”是 while 引导让步状语从句。全句译为“电涡流传动尽管从技术的角度来说属于电气传动,然而在功能上很像流体动力传动和液体黏性传动,它起着原动力和被驱动负载之间的耦合器的作用。”

[2] In addition, by reducing, or often eliminating, the need for wear-prone mechanical components, electrical drives foster increased overall system reliability, as well as lower maintenance costs. “by reducing … components”为介词短语做状语,“increased overall system reliability 和 lower maintenance costs”为并列谓语。全句译为“另外,通过减少或者省去易于磨损的机械部件,电气传动提供了更强的系统可靠性和更低的维护成本。”

[3] The filtering section and higher level of control modulation account for the PWM drive's improved performance compared with a common SCR drive. 句中“The filtering … control modulation”为主语,account for 的意思为“对……负责”,“compared with a common SCR drive”是过去分词做状语,限定“performance”。全句译为“和可控硅传动比较,滤波部分和更高一层的调制控制是为提高 PWM 传动的性能。”

[4] Plus, microprocessor advancements have enabled the creation of so-called vector drives, which provide greatly enhance response, operation down to zero speed and positioning accuracy. 句中“which provide…”为定语从句,修饰“vector drives”。从句中有三个并列宾语“greatly enhance response”、“operation down to zero speed”和“positioning accuracy。”全句译为“更有,微处理器的进步使得所谓矢量传动的建立成为可能,这种传动系统使得系统响应能力大大增强,且实现了低至零速的控制及精确定位。”

[5] Vector drives, especially when combined with feedback devices such as tachometers, encoders and resolvers in a closed-loop system, are continuing to replace DC drives in demanding applications. 句中“when combined with feedback devices…”为省略句,省略了主语和谓语。全句译为“特别是当矢量传动和反馈设备结合例如转速测量计、编码器和旋转变压器一起构成闭环系统时,它正在高要求的应用场合下不断地替代直流传动。”

[6] Though originally developed for smaller-horsepower applications, PWM is now used in drives of hundreds or even thousands of horsepower — as well as remaining the staple technology in the vast majority of small integral and fractional horsepower “micro” and “sub-micro” AC drives. 句中“Though originally developed for smaller-horsepower applications”为过去分词短语做状语,全句译为“尽管最初研究它是用于小功率应用的,但现在,除了在数量庞大的小马力和分马力的微型和亚微型交流传动中作为主要技术手段外,PWM 已经用在几百甚至数千马力的传动中了。”

[7] A big advantage to these devices is their fast switching speed resulting in higher pulse or carrier frequency, which minimizes motor noise. “resulting in higher pulse or carrier frequency”为现在分词短语做结果状语,“which minimizes motor noise”则是非限定性定语从句。全句译为“这些器件的优点是开关速度高,这样可以产生更高的脉冲频率或者脉冲载波频率,它可使电动机的噪声最小化。”

6.3.4 Reference Translation

调速传动

根据定义,任何类型的速度可调的传动都会提供一种易于改变转速的方法以适应运行的需要。这样的传动存在有机械的、流体的和电气的三种类型。

最常见的机械方式采用三角皮带轮或者链轮,按照事先设定好的、可选的比例来调节速度,例如 $2:1$ 、 $4:1$ 、 $8:1$ 等。更为复杂的牵引传动方案则允许增量式的速度调节,它是通过改变金属盘或者球和锥体之间的接触点来实现调速的。

可调速的液压传动可以实现平滑无级的速度控制。它主要有三种类型。静液压传动采用电动机或者内燃机作为原动力,与液压泵相结合再来驱动液压马达。流体动力传动和液体黏性传动直接将输入和输出轴相耦合。流体动力传动是通过改变作为输入输出耦合器的涡流量的方法来调节速度的。液体黏性传动也称为油膜剪切传动,是通过控制油膜的厚度来调节速度的,因此在旋转的金属盘间会产生滑移。

电涡流传动尽管从技术的角度来说属于电气传动,然而在功能上很像流体动力传动和液体黏性传动,它起着原动力和被驱动负载之间的耦合器的作用。在电涡流传动中,耦合由原电磁场和次电磁场组成,其中磁场是由感应电涡流产生的。通过允许的磁场间的电磁滑移量来控制传动的速度。

在大多数工业应用中,机械的、流体的或是电涡流传动多用于恒速电机中。而固态电

气传动(也称为电气传动)建立的是可调速的电机系统,它允许转速从零到基速以上可调。控制电机转速有多种好处,包括通过消除在机械转速控制设备中的损耗而提高效率。另外,通过减少或者省去易于磨损的机械部件,电气传动具有更强的系统可靠性和更低的维护成本。由于多方面的原因,电气传动成为可调速的传动中增长最快的类型。

根据被控的电机的不同(直流和交流),可以有两种基本的传动类型。直流传动通过改变电枢电压(有时也称为励磁电压)来控制直流电机的转速。交流传动通过改变供给电动机的频率和电压来控制交流电动机的转速。

直流传动

直流传动易于使用,技术上直截了当。它是通过将电网的交流电压整流为直流电压,然后把可调的直流电压供给直流电机的。对于永磁直流电机来说,仅控制电枢电压。提供的电压越高,电枢就转动越快。对于绕线磁极式电动机来说,电压必须同时施加于电枢和磁场上。在工业中,下列三种直流传动最为常见,如图 6.3.1 所示:

直流可控硅传动:这是以可控硅整流器(也称为晶闸管)的名字来命名的,它用来实现交流电压到直流电压的转换。这类传动价格便宜且易于使用,封装样式很多,有单向或者可反向的类型。

回馈式可控硅传动:也称为四象限传动,这种方式可以使直流电机提供驱动和制动两种转矩。电动机在制动时产生的电能将反馈回电网,而不会浪费掉。

脉宽调制型直流传动:缩写为 PWM,也被更一般地称为晶体管化直流传动,这类直流传动可以提供更平滑的速度控制,效率高且电动机发热量小。不同于可控硅传动, PWM 型传动有三个组成部分。第一部分是交流到直流的变换,第二部分是滤波并调整为固定的直流电压,第三部分是通过产生一个可调节直流脉冲序列以控制平均电压。滤波部分和更高级的调制控制使得 PWM 传动性能优于可控硅。

交流传动

交流传动系统的开始部分和直流传动系统非常相似。首先交流电网的电压整流以产生直流电压。但由于这里使用的是交流电动机,所以直流电压必须反转,或者称为逆变,变成一种频率可调的交流电压。交流传动的逆变部分可以完成这部分功能。过去这是采用可控硅来实现的。而现代交流传动采用若干晶体管来实现直流到频率可调的交流的逆变过程。图 6.3.2 给出了一个示例。

按一定的频率和电压,将这样合成的交流电提供给交流电动机,得到所需要的转速。例如可以合成出和美国电网标准一样的 60 Hz 的频率,使电动机达到 100% 的额定转速。低频率对应低转速,高频率对应高转速。按照这种方式,交流传动能驱动的电动机的转速可以通过将频率从 9 Hz 调到 120 Hz,而实现 15% 的额定转速到 200% 的额定转速。

目前,许多工业中交流传动正在成为主流选择。使用简单且耐用的三相感应电机意味着交流传动系统几乎是所有系统中最可靠且维护成本最少的。更有,微处理器的进步使得所谓矢量传动的建立成为可能,这种传动系统使得系统响应能力大大增强,且实现了低至零速的控制及精确定位。特别是当矢量传动和反馈设备结合例如转速测量计、编码

器和旋转变压器一起构成闭环系统时,它正在高要求的应用场合下不断地替代直流传动。图 6.3.3 给出了一个示例。

到目前为止,应用最多的交流传动为脉宽调制型。尽管最开始研究它是用于小功率应用的,但现在,除了在数量庞大的小马力和分马力的微型和亚微型交流传动中继续作为主要的技术手段外,PWM 已经用在几百甚至数千马力的传动中了,如图 6.3.4 所示。

脉宽调制关系到逆变器通过改变脉冲电压的宽度和极性来改变输出电压的能力。电压和频率是通过电压脉冲序列来合成所需要的电压和频率的。而这些是由微处理器对工作在开关状态的电力电子器件发出指令来完成的。现在通常使用 IGBT 或称为绝缘栅极双极型晶体管来作为开关器件。这些器件的优点是开关速度高,这样可以产生更高的脉冲频率或者脉冲载波频率,它可使电动机的噪声最小化。

6.3.5 Reading Materials

AC Drive Application Factors

As PWM AC drives have continued to increase in popularity, drives manufacturers have spent considerable research and development effort to build in programmable acceleration and deceleration ramps, a variety of speed presets, diagnostic abilities, and other software features. Operator interfaces have also been improved with some drives incorporating “plain-English” readouts to aid set-up and operation. Plus, an array of input and output connections, plug-in programming modules, and off-line programming tools allow multiple drive set-ups to be installed and maintained in a fraction of the time spent previously. All these features have simplified drive applications. However, several basic points must be considered:

Torque(转矩)

This is the most critical application factor. All torque requirements must be assessed, including starting, running, accelerating and decelerating and, if required, holding torque. These values will help determine what current capacity the drive must have in order for the motor to provide the torque required. Usually, the main constraint is starting torque, which relates to the drive’s current overload capacity. (Many drives also provide a starting torque boost by increasing voltage at lower frequencies.)

Perhaps the overriding question, however, is whether the application is variable torque or constant torque. Most variable torque applications fall into one of two categories-air moving or liquid moving-and involve centrifugal pumps and fans. The torque required in these applications decreases as the motor RPM decreases. Therefore, drives for variable torque loads require little overload capacity. Constant torque applications, including conveyors, positive displacement pumps, extruders, mixers or other “machinery” require the same torque regardless of operating speed, plus extra

torque to get started. Here, high overload capacity is required.

Smaller-horsepower drives are often built to handle either application. Typically, only a programming change is required to optimize efficiency (variable volts-to-hertz ratio for variable torque loads, constant volts-to-hertz ratio for constant torque loads). Larger horsepower drives are usually built specifically for either variable or constant torque applications.

Speed

As mentioned, AC drives provide an extremely wide speed range. In addition, they can provide multiple means to control this speed. Many drives, for example, include a wide selection of preset speeds, which can make set-up easier. Similarly, a range of acceleration and deceleration speed “ramps” are provided. Slip compensation, which maintains constant speed with a changing load, is another feature that can be helpful. In addition, many drives have programmable “skip frequencies.” Particularly with fans or pumps, there may be specific speeds at which vibration takes place. By programming the drive to avoid these corresponding frequencies, the vibration can be minimized. Another control function, common with fans, is the ability for the drive to start into a load already in motion—often called a rolling start or spinning start. If required, be sure your drive allows this or you will face overcurrent tripping.

Current

The current a motor requires to provide needed torque (see previous discussion of torque) is the basis for sizing a drive. Horsepower ratings, while listed by drives manufacturers as a guide to the maximum motor size under most applications, are less precise. Especially for demanding constant torque applications, the appropriate drive may, in fact, be “oversized” relative to the motor. As a rule, general-purpose constant torque drives have an overload current capacity of approximately 150% for one minute, based on nominal output. If an application exceeds these limits, a larger drive should be specified.

Power Supply

Drives tolerate line-voltage fluctuations of 10-15% before tripping and are sensitive to power interruptions. Some drives have “ride-through” capacity of only a second or two before a fault is triggered, shutting down the drive. Drives are sometimes programmed for multiple automatic restart attempts. For safety, plant personnel must be aware of this. Manual restart may be preferred.

Most drives require three-phase input. Smaller drives may be available for single-phase input. In either case, the motor itself must be three-phase.

Drives, like any power conversion device, create certain power disturbances (called

“noise” or “harmonic distortion”) that are reflected back into the power system to which they are connected. These disturbances rarely affect the drive itself but can affect other electrically sensitive components.

Control Complexity: Even small, low-cost AC drives are now being produced with impressive features, including an array of programmable functions and extensive input and output capability for integration with other components and control systems. Additional features may be offered as options. Vector drives, as indicated previously, are one example of enhanced control capability for specialized applications.

In addition, nearly all drives provide some measure of fault logging and diagnostic capability. Some are extensive, and the easiest to use display the information in words and phrases rather than simply numerical codes.

Environmental Factors

The enemies of electronic components are well-known. Heat, moisture, vibration and dirt are chief among them and obviously should be mitigated. Drives are rated for operation in specific maximum and minimum ambient temperatures. If the maximum ambient is exceeded, extra cooling must be provided, or the drive may have to be oversized. High altitudes, where thinner air limits cooling effectiveness, call for special consideration. Ambient temperatures too low can allow condensation. In these cases, or where humidity is generally high, a space heater may be needed.

6.4 Power Semiconductor Devices

6.4.1 Text

The modern age of power electronics began with the introduction of thyristors in the late 1950s. Now there are several types of power devices available for high-power and high-frequency applications. The most notable power devices are gate turn-off thyristors, power Darlington transistors, power MOSFETs, and insulated-gate bipolar transistors (IGBTs). Power semiconductor devices are the most important functional elements in all power conversion applications. The power devices are mainly used as switches to convert power from one form to another. They are used in motor control systems, uninterrupted power supplies, high-voltage dc transmission, power supplies, induction heating, and in many other power conversion applications. A review of the basic characteristics of these power devices is presented in this section.

Thyristor

The thyristor, also called a silicon-controlled rectifier (SCR), is basically a four-layer three-junction pn device. It has three terminals; anode, cathode, and gate. The

device is turned on by applying a short pulse across the gate and cathode. Once the device turns on, the gate loses its control to turn off the device. The turn-off is achieved by applying a reverse voltage across the anode and cathode. The thyristor symbol and its volt-ampere characteristics are shown in Figure 6. 4. 1. There are basically two classifications of thyristors; converter grade and inverter grade. The difference between a converter-grade and an inverter-grade thyristor is the low turn-off time (on the order of a few microseconds) for the latter. The converter-grade thyristors are slow type and are used in natural commutation (or phase-controlled) applications. Inverter-grade thyristors are used in forced commutation applications such as dc-dc choppers and dc-ac inverters. The inverter-grade thyristors are turned off by forcing the current to zero using an external commutation circuit. This requires additional commutating components, thus resulting in additional losses in the inverter.

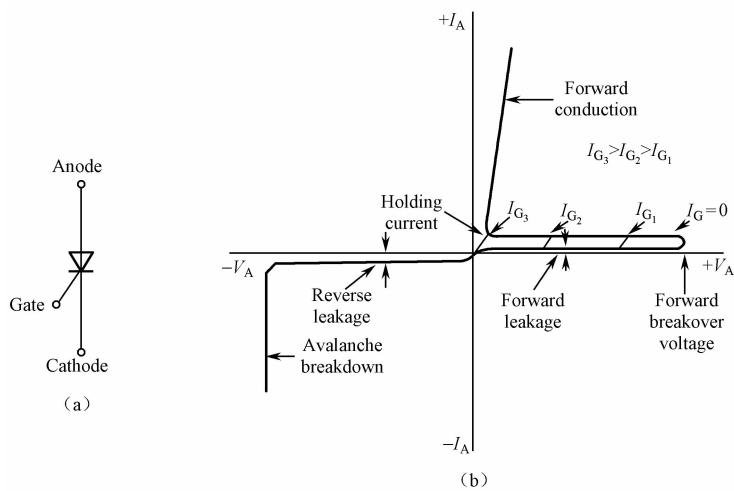


Figure 6. 4. 1 (a) Thyristor symbol and (b) volt-ampere characteristics.

Thyristors are highly rugged devices in terms of transient currents, di/dt , and dv/dt capability. The forward voltage drop in thyristors is about 1. 5 to 2 V, and even at higher currents of the order of 1000 A, it seldom exceeds 3 V. While the forward voltage determines the on-state power loss of the device at any given current, the switching power loss becomes a dominating factor affecting the device junction temperature at high operating frequencies.^[1] Because of this, the maximum switching frequencies possible using thyristors are limited in comparison with other power devices considered in this section.

Thyristors have I^2t withstand capability and can be protected by fuses. The nonrepetitive surge current capability for thyristors is about 10 times their rated root

mean square (rms) current. They must be protected by snubber networks for dv/dt and di/dt effects. If the specified dv/dt is exceeded, thyristors may start conducting without applying a gate pulse. In dc-to-ac conversion applications it is necessary to use an antiparallel diode of similar rating across each main thyristor. Thyristors are available up to 6000 V, 3500 A.

Power MOSFET

Power MOSFETs are marketed by different manufacturers with differences in internal geometry and with different names such as MegaMOS, HEXFET, SIPMOS, and TMOS. They have unique features that make them potentially attractive for switching applications. They are essentially voltage-driven rather than current-driven devices, unlike bipolar transistors.^[2]

The gate of a MOSFET is isolated electrically from the source by a layer of silicon oxide. The gate draws only a minute leakage current of the order of nanoamperes. Hence the gate drive circuit is simple and power loss in the gate control circuit is practically negligible. Although in steady state the gate draws virtually no current, this is not so under transient conditions. The gate-to-source and gate-to-drain capacitances have to be charged and discharged appropriately to obtain the desired switching speed, and the drive circuit must have a sufficiently low output impedance to supply the required charging and discharging currents.^[3] The circuit symbol of a power MOSFET is shown in Figure 6.4.2.

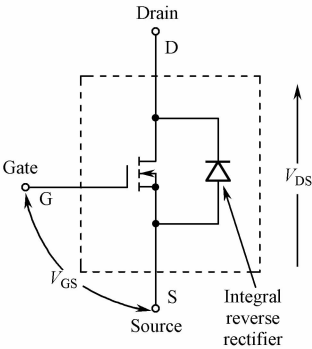


Figure 6.4.2 Power MOSFET circuit symbol.

Power MOSFETs are majority carrier devices, and there is no minority carrier storage time. Hence they have exceptionally fast rise and fall times. They are essentially resistive devices when turned on, while bipolar transistors present a more or less constant $V_{CE(sat)}$ over the normal operating range. Power dissipation in MOSFETs is $I_d^2 R_{DS(on)}$, and in bipolar it is $I_C V_{CE(sat)}$. At low currents, therefore, a power MOSFET may have a lower conduction loss than a comparable bipolar device, but at higher currents, the conduction loss will exceed that of bipolar. Also, the $R_{DS(on)}$ increases with temperature.

An important feature of a power MOSFET is the absence of a secondary breakdown effect, which is present in a bipolar transistor, and as a result, it has an extremely rugged switching performance. In MOSFETs, $R_{DS(on)}$ increases with temperature, and

thus the current is automatically diverted away from the hot spot. The drain body junction appears as an antiparallel diode between source and drain. Thus power MOSFETs will not support voltage in the reverse direction. Although this inverse diode is relatively fast, it is slow by comparison with the MOSFET. Recent devices have the diode recovery time as low as 100 ns. Since MOSFETs cannot be protected by fuses, an electronic protection technique has to be used.

With the advancement in MOS technology, ruggedized MOSFETs are replacing the conventional MOSFETs. The need to ruggedize power MOSFETs is related to device reliability. If a MOSFET is operating within its specification range at all times, its chances for failing catastrophically are minimal. However, if its absolute maximum rating is exceeded, failure probability increases dramatically. Under actual operating conditions, a MOSFET may be subjected to transients—either externally from the power bus supplying the circuit or from the circuit itself due, for example, to inductive kicks going beyond the absolute maximum ratings. Such conditions are likely in almost every application, and in most cases are beyond a designer's control.^[4] Rugged devices are made to be more tolerant for over-voltage transients. Ruggedness is the ability of a MOSFET to operate in an environment of dynamic electrical stresses, without activating any of the parasitic bipolar junction transistors. The rugged device can withstand higher levels of diode recovery dv/dt and static dv/dt .

Insulated-Gate Bipolar Transistor (IGBT)

The IGBT has the high input impedance and high-speed characteristics of a MOSFET with the conductivity characteristic (low saturation voltage) of a bipolar transistor. The IGBT is turned on by applying a positive voltage between the gate and emitter and, as in the MOSFET, it is turned off by making the gate signal zero or slightly negative. The IGBT has a much lower voltage drop than a MOSFET of similar ratings. The structure of an IGBT is more like a thyristor and MOSFET. For a given IGBT, there is a critical value of collector current that will cause a large enough voltage drop to activate the thyristor. Hence, the device manufacturer specifies the peak allowable collector current that can flow without latch-up occurring. There is also a corresponding gate source voltage that permits this current to flow that should not be exceeded.

Like the power MOSFET, the IGBT does not exhibit the secondary breakdown phenomenon common to bipolar transistors. However, care should be taken not to exceed the maximum power dissipation and specified maximum junction temperature of the device under all conditions for guaranteed reliable operation. The on-state voltage of the IGBT is heavily dependent on the gate voltage. To obtain a low on-state voltage, a

sufficiently high gate voltage must be applied.

In general, IGBTs can be classified as punch-through (PT) and nonpunch-through (NPT) structures, as shown in Figure 6. 4. 3. In the PT IGBT, an N^+ buffer layer is normally introduced between the P^+ substrate and the N^- epitaxial layer, so that the whole N^- drift region is depleted when the device is blocking the off-state voltage, and the electrical field shape inside the N^- drift region is close to a rectangular shape. Because a shorter N^- region can be used in the punch-through IGBT, a better trade-off between the forward voltage drop and turn-off time can be achieved. PT IGBTs are available up to about 1200 V.

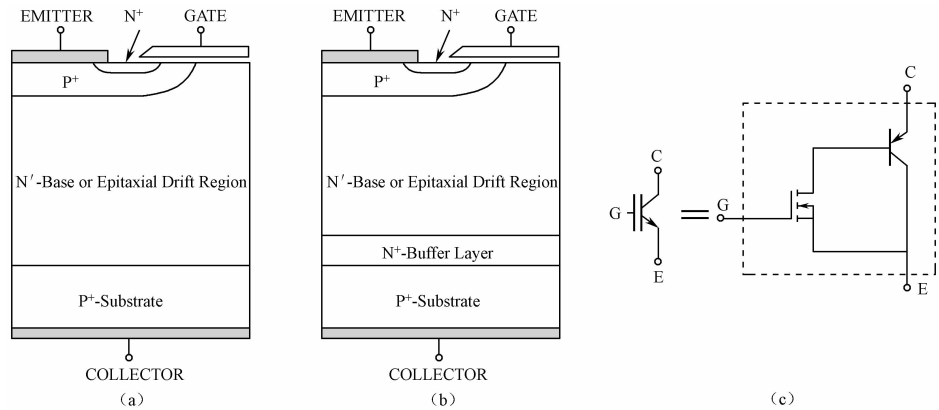


Figure 6. 4. 3 (a) Nonpunch-through IGBT, (b) Punch-through IGBT, (c) IGBT equivalent circuit.

High voltage IGBTs are realized through nonpunch-through process. The devices are built on a N^- wafer substrate which serves as the N^- base drift region. Experimental NPT IGBTs of up to about 4 kV have been reported in the literature. NPT IGBTs are more robust than PT IGBTs particularly under short circuit conditions. But NPT IGBTs have a higher forward voltage drop than the PT IGBTs.

The PT IGBTs cannot be as easily paralleled as MOSFETs. The factors that inhibit current sharing of parallel-connected IGBTs are (1) on-state current unbalance, caused by $V_{CE(sat)}$ distribution and main circuit wiring resistance distribution, and (2) current unbalance at turn-on and turn-off, caused by the switching time difference of the parallel connected devices and circuit wiring inductance distribution. The NPT IGBTs can be paralleled because of their positive temperature coefficient property.

6. 4. 2 Specialized English Words

thyristor	晶闸管	anode	阳极
Darlington transistors	达林顿管	cathode	阴极

gate 门极,栅极	electrical stresses 电应力
uninterrupted power supplies 不间断电源	parasitic 寄生的
natural commutation 自然换相	inductive kick 感性冲击
phase-controlled 相控	conductivity 导电率
on-state 导通状态	latch-up 擎住效应
root mean square (rms) 均方根	punchthrough(PT) 穿透型
antiparalled 反并联的	substrate 衬底,基质
ac regulators 交流调压器	epitaxial layer 外延层
solid-state ac relays 固态交流继电器	drift region 漂移区
snubber circuit 缓冲电路	deplete 耗尽
carriers 载流子	rectangular 长方形,矩形
exceptionally 指数的	wafer 硅片,晶片
catastrophically 灾难的	inhibit 约束,抑制
critical value 临界值	holding current 维持电流
saturation voltage 饱和电压	avalanche breakdown 雪崩击穿

6.4.3 Notes

[1] While the forward voltage determines the on-state power loss of the device at any given current, the switching power loss becomes a dominating factor affecting the device junction temperature at high operating frequencies. 全句由两个并列句组成,构成转折关系,“affecting”为现在分词做定语,修饰“dominating factor”。全句译为“在低频时,对于任意的给定电流,正向压降是影响通态功率损耗的主要因素,而工作在高频时,开关损耗则是影响器件结温的主要因素。”

[2] They are essentially voltage-driven rather than current-driven devices, unlike bipolar transistors. “Rather than”译为“而不是……”。全句译为“不同于双极型晶体管是电流驱动器件,它们基本上为电压驱动型。”

[3] The gate-to-source and gate-to-drain capacitances have to be charged and discharged appropriately to obtain the desired switching speed, and the drive circuit must have a sufficiently low output impedance to supply the required charging and discharging currents. 句中“and”为并列连词,全句由两个并列的句子组成。全句译为“为了得到理想的开关速度,栅漏间电容和栅源间电容必须适当地充电放电,而驱动电路阻抗必须足够低,以便能提供所需的充电和放电电流。”

[4] Under actual operating conditions, a MOSFET may be subjected to transients—either externally from the power bus supplying the circuit or from the circuit itself due, for example, to inductive kicks going beyond the absolute maximum ratings. 本句虽然结构上属于简单句,但成分较复杂。两个并列的“from”介词短语是“itself due, for

example, to inductive kicks going beyond the absolute maximum ratings” 的状语, “supplying the circuit” 为现在分词短语做 “the power bus” 的后置定语, “due, for example, to ...” 则表示原因。全句译为 “在实际运行条件中, 场效应管也许会受到比如给电路供电的外部电源母线或者电路本身的感性冲击产生的过渡过程的限制, 感性冲击造成的电流电压有可能超过最大额定值。”

6.4.4 Reference Translation

电力半导体器件

现代电力电子学是以 20 世纪 50 年代末期引入晶闸管开始的。目前有多种电力器件可以用于大功率和高频的应用场合。最值得注意的电力器件有门极可关断晶闸管、电力达林顿晶体管、电力场效应管和绝缘栅极型晶体管(IGBTs)。电力半导体器件在所有的电力转换应用中是最重要的功能性元件。电力器件主要被用做开关, 将电力从一种形式转换到另外一种形式。它们被广泛应用于电动机控制系统、不间断电源、高压直流输电、电源、感应加热以及许多其他的电力变换领域。在本节中, 将对这些电力器件的基本特性进行回顾。

晶闸管

晶闸管, 也称可控硅, 是一种四层三个 PN 节的 PNP 型器件。它有三个端: 阳极、阴极和门极。通过在门极和阴极之间施加窄脉冲就可以使其导通。器件一旦导通, 门极将无法使其关断。在阳极和阴极直接施加反相电压可以使其关断。晶闸管的符号和它的伏安特性如图 6.4.1 所示。晶闸管大体上分为两类: 变换器级和逆变级。区别在于后者比前者的关断时间要小(数微秒的数量级)。变换器级的晶闸管属于慢速类型, 用于自然换向(或相控)应用中。变频级的晶闸管用于强迫关断的应用中, 例如斩波器和直流-交流的逆变器。逆变级晶闸管的关断采用的是用外部换相电路使其电流降到零的方法。这样就需要额外的换向部件, 因此会在变频器中产生额外的损耗。

从暂态电流、 di/dt 和 dv/dt 的指标来看, 晶闸管是一种很耐用的器件。晶闸管的正向电压降大约为 1.5~2 V, 即使电流等级为 1000 A, 正向电压降也很少超过 3 V。在低频时, 对于任意的给定电流, 正向压降是影响通态功率损耗的主要因素, 而工作在高频时, 开关损耗则是影响器件结温的主要因素。正因如此, 相对本节讨论的其他电力器件, 晶闸管可以使用的最大工作频率是很有限的。

晶闸管具有 I^2t 承受能力, 可以用熔断丝来保护。对非重复性浪涌电流的耐受力大约是其有效值(rms)的 10 倍。由于有 dv/dt 和 di/dt 影响, 必须采用缓冲电路来保护。如果超过规定的 dv/dt , 晶闸管也许会在没有施加门极电流的情况下导通。在逆变的应用中, 有必要给每个主晶闸管反并联一个与其额定值相近的二极管。晶闸管额定值最高可以到 6000 V, 3500 A。

电力场效应管

市场上有不同的制造商生产的不同内部结构和不同名字的电力场效应管, 例如

MegaMOS、HEXFET、SIPMOS 和 TMOS。它们独特的性能使它们在开关方面的应用非常具有吸引力。不同于双极型晶体管是电流驱动器件,它们基本上为电压驱动型。

场效应管的栅极被一层氧化硅从电气上隔离。栅极仅流过纳秒级的非常微小的电流。因此栅极驱动电流简单,而且栅极控制电路的功率损耗在实际中可以忽略。尽管在稳态时栅极几乎没有电流,但是在过渡过程中并非如此。为了得到理想的开关速度,栅漏间电容和栅源间电容必须适当地充电放电,而驱动电路阻抗必须足够低,以便能提供所需的充电和放电电流。电力场效应管的电路符号如图 6.4.2 所示。

电力场效应管是一种多数载流子器件,无少数载流子存储时间。因此它们的上升和下降时间都非常快。当它们导通时基本上属于电阻性的器件,而双极型晶体管在正常工作范围内的 $V_{CE(sat)}$ 基本是恒值。在场效应管中的功率损耗为 $I_d^2 R_{DS(on)}$,在双极型晶体管中为 $I_C V_{CE(sat)}$ 。如果电流很小,电流场效应管比双极型晶体管的导通损耗小,但是在电流中,导通损耗将超过双极型晶体管。同样, $R_{DS(on)}$ 也随着温度的升高而增加。

电力场效应管的一个重要特点是没有二次击穿效应,而双极型晶体管有,因此电力场效应管具有非常耐用的开关特性。在场效应管中, $R_{DS(on)}$ 随着温度的升高而升高,因此电流自动离开热的区域。体漏极 PN 结看起来如同漏源之间的一个反并联的二极管。因此电力场效应管无法承受反向电压。尽管反向二极管的速度相对快,但是和场效应管的速度比起来还是慢。近来器件可以让反向二极管做到具有 100 ns 的反向恢复时间。由于场效应管无法用熔断器保护,因此必须采用电子保护措施。

随着金属氧化半导体技术的进步,耐用的场效应管正在逐步替代传统的场效应管。电力场效应管的耐用性主要涉及器件的可靠性。如果场效应管在所有时段都工作在其规定参数范围内,则它彻底损坏的可能性将降到最低。然而,一旦超过场效应管最大额定值,其损坏的可能性将大大增加。在实际运行条件中,场效应管也许会受到比如给电路供电的外部电源母线或者电路本身的感性冲击产生的过渡过程的限制,感性冲击造成的电流电压有可能超过最大额定值。这种情况在所有的应用中几乎都存在,而且大多数情况下都是设计者无法控制的。耐用的器件对过电压的过渡过程具有更好的耐受能力。耐用性是场效应管在动态电应力环境中能正常工作而不会激活任何寄生双极结型晶体管的一种能力。耐用器件的二极管能承受更高的二极管恢复 dv/dt 和静态的 dv/dt 。

绝缘栅极双极型晶体管(IGBT)

IGBT 具有 MOSFET 的高输入阻抗和高速特性,以及双极型晶体管的高导电率特性(低饱和电压)。和 MOSFET 一样,在 IGBT 的栅极和发射极施加正向电压可以使其导通,使其门极信号为零或者施以小的反向电压可以将它关断。在相似的额定值下,IGBT 具有比 MOSFET 小得多的电压降。IGBT 的结构更像晶闸管和 MOSFET。IGBT 的集电极电流的临界值会引起足够大的电压降而激活晶闸管。因此器件制造商规定允许的集电极峰值电流不会产生擎住效应。对应有一个栅源电压允许该电流通过但也不能超过。

同电力 MOSFET 一样,IGBT 也没有双极晶体管常见的二次击穿现象。然而,为了保证可靠运行,在任何情况下都不允许超过最大耗散功率和器件的规定的最大结温。

IGBT 的通态电压主要与栅极电压有关。为了获得低通态电压,需要施加足够大的栅极电压。

一般来说,IGBT 可以分为穿透式(PT)和非穿透式(NPT)两种,如图 6.4.3 所示。在穿透式 IGBT 中, N^+ 缓冲层通常位于 P^+ 衬底和 N^- 外延层之间,因此当器件处于断态时,整个 N^- 漂移区耗尽,其的电场形状接近长方形。由于在穿透型 IGBT 中可使用较短的 N^- 区,这样可在正向压降和关断时间之间找到一个更好的折中点。穿透型 IGBT 额定值最大可以达到 1200 V。

高压 IGBT 用非穿透型工艺制成。这种器件制作在 N^- 型的硅片基层上,该基层作用为 N^- 基区漂移区。有文献报道,实验的非穿透型 IGBT 可以到达 4 kV。特别是在短路情况下,非穿透型 IGBT 比穿透型的更耐用。但是非穿透型 IGBT 的正向压降比穿透型的高。

穿透型 IGBT 不能像 MOSFET 一样易于并联。制约 IGBT 并联的因素为:(1)通态下由 $V_{CE(sat)}$ 分布和主电路导线电阻分布引起的电流不平衡,(2)由并联器件的开关时间的差异和电路导线电感分布所引起的在导通和关断时刻的电流不平衡。非穿透型 IGBT 由于有正温度系数的特性,所以易于并联。

6.4.5 Reading Materials

Power Transistor

Power transistors are used in applications ranging from a few to several hundred kilowatts and switching frequencies up to about 10 kHz. Power transistors used in power conversion applications are generally npn type. The power transistor is turned on by supplying sufficient base current, and this base drive has to be maintained throughout its conduction period. It is turned off by removing the base drive and making the base voltage slightly negative (within $-V_{BE(max)}$). The saturation (饱和) voltage of the device is normally 0.5 to 2.5 V and increases as the current increases. Hence the on-state losses increase more than proportionately with current. The transistor off-state losses are much lower than the on-state losses because the leakage current of the device is of the order of a few milliamperes. Because of relatively larger switching times, the switching loss significantly increases with switching frequency. Power transistors can block only forward voltages. The reverse peak voltage rating of these devices is as low as 5 to 10 V.

Power transistors do not have I^2t withstand capability. In other words, they can absorb only very little energy before breakdown. Therefore, they cannot be protected by semiconductor fuses, and thus an electronic protection method has to be used.

To eliminate high base current requirements, Darlington configurations are commonly used. They are available in monolithic or in isolated packages. The basic

Darlington configuration is shown schematically in Figure 6.4.4. The Darlington configuration presents a specific advantage in that it can considerably increase the current switched by the transistor for a given base drive. The $V_{CE(sat)}$ for the Darlington is generally more than that of a single transistor of similar rating with corresponding increase in on-state power loss. During switching, the reverse-biased collector junction may show hot spot breakdown effects that are specified by reverse-bias safe operating area (RBSOA) and forward bias safe operating area (FBSOA). Modern devices with highly interdigitated emitter base geometry force more uniform current distribution and therefore considerably improve second breakdown effects. Normally, a well-designed switching aid network constrains the device operation well within the SOAs.

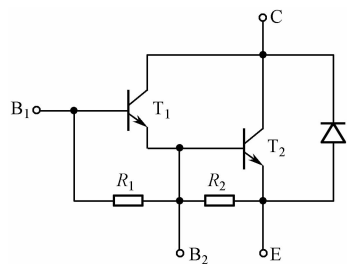


Figure 6.4.4 A two-stage Darlington transistor with bypass diode.

6.5 DC Power Supply

6.5.1 Text

Power supplies are used in many industrial and aerospace applications and also in consumer products. Some of the requirements of power supplies are small size, light weight, low cost, and high power conversion efficiency. In addition to these, some power supplies require the following: electrical isolation between the source and load, low harmonic distortion for the input and output waveforms, and high power factor (PF) if the source is ac voltage. Some special power supplies require controlled direction of power flow.

Basically two types of power supplies are required: dc power supplies and ac power supplies. The output of dc power supplies is regulated or controllable dc, whereas the output for ac power supplies is ac. The input to these power supplies can be ac or dc.

If an ac source is used, then ac-to-dc converters explained in Section 30.2 can be used. In these converters, electrical isolation can only be provided by bulky line frequency transformers. The ac source can be rectified with a diode rectifier to get an uncontrolled dc, and then a dc-to-dc converter can be used to get a controlled dc output.

Electrical isolation between the input source and the output load can be provided in the dc-to-dc converter using a high-frequency (HF) transformer. Such HF transformers have small size, light weight, and low cost compared to bulky line frequency transformers. Whether the input source is dc (e. g. , battery) or ac, dc-to-dc converters form an important part of dc power supplies, and they are explained in this subsection.

DC power supplies can be broadly classified as linear and switching power supplies.

A linear power supply is the oldest and simplest type of power supply. The output voltage is regulated by dropping the extra input voltage across a series transistor (therefore, also referred to as a series regulator). They have very small output ripple, theoretically zero noise, large hold-up time (typically 1-2 ms), and fast response. Linear power supplies have the following disadvantages: very low efficiency, electrical isolation can only be on 60 Hz ac side, larger volume and weight, and, in general, only a single output possible. However, they are still used in very small regulated power supplies and in some special applications (e. g. , magnet power supplies). Three terminal linear regulator integrated circuits (ICs) are readily available (e. g. , μ A7815 has +15 V, 1 A output), are easy to use, and have built-in load short-circuit protection.

Switching power supplies use power semiconductor switches in the on and off switching states resulting in high efficiency, small size, and light weight. With the availability of fast switching devices, HF magnetics and capacitors, and high-speed control ICs, switching power supplies have become very popular. They can be further classified as pulsewidth-modulated (PWM) converters and resonant converters, and they are explained below.

Pulsewidth-Modulated Converters

These converters employ square-wave pulsewidth modulation to achieve voltage regulation. The average output voltage is varied by varying the duty cycle of the power semiconductor switch. The voltage waveform across the switch and at the output are square wave in nature and they generally result in higher switching losses when the switching frequency is increased. Also, the switching stresses are high with the generation of large electromagnetic interference (EMI), which is difficult to filter^[1]. However, these converters are easy to control, well understood, and have wide load control range. The methods of control of PWM converters are discussed next.

The Methods of Control. The PWM converters operate with a fixed-frequency, variable duty cycle. Depending on the duty cycle, they can operate in either continuous current mode (CCM) or discontinuous current mode (DCM). If the current through the output inductor never reaches zero, then the converter operates in CCM; otherwise

DCM occurs.

The three possible control methods are briefly explained below.

1. *Direct duty cycle control*. It is the simplest control method. A fixed-frequency ramp is compared with the control voltage (Figure 6. 5. 1(a)) to obtain a variable duty cycle base drive signal for the transistor. This is the simplest method of control. Disadvantages of this method are (a) provides no voltage feedforward to anticipate the effects of input voltage changes, slow response to sudden input changes, poor audio susceptibility, poor open-loop line regulation, requiring higher loop gain to achieve specifications; (b) poor dynamic response.

2. *Voltage feedforward control*. In this case the ramp amplitude varies in direct proportion to the input voltage (Figure 6. 5. 1(b)). (The open-loop regulation is very good, and the problems in 1(a) above are corrected.)

3. *Current mode control*. In this method, a second inner control loop compares the peak inductor current with the control voltage which provides improved open-loop line regulation (Figure 6. 5. 1(c))^[2]. All the problems of the direct duty cycle control method 1 above are corrected with this method. An additional advantage of this method is that the two-pole second-order filter is reduced to a single-pole (the filter capacitor) first-order filter, resulting in simpler compensation networks^[3].

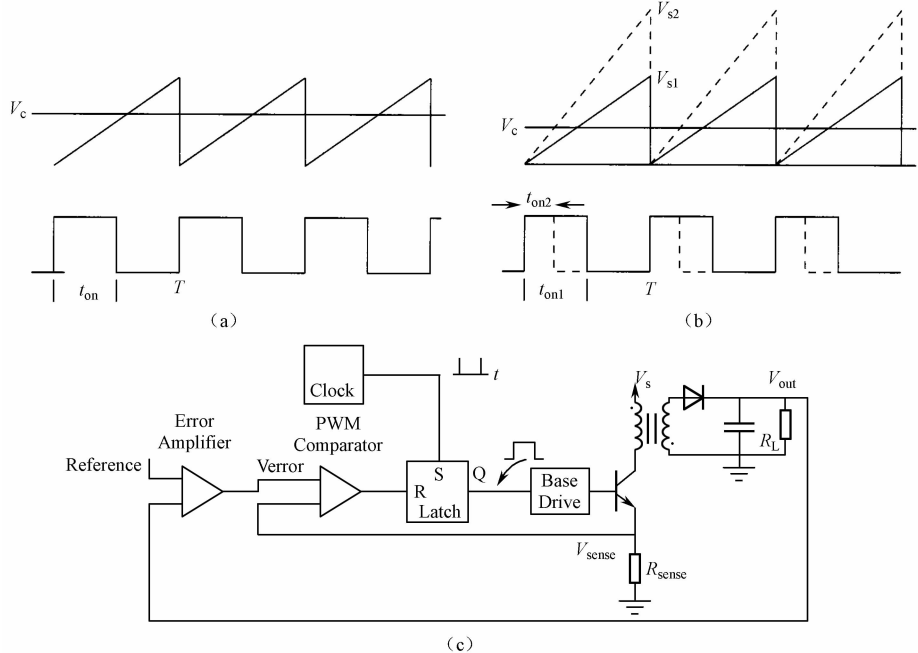


Fig 6. 5. 1 PWM converter control methods: (a) direct duty cycle control; (b) voltage feedforward control; (c) current mode control (illustrated for flyback converter).

The above control methods can be used in all the PWM converter configurations explained below.

PWM converters can be classified as single-ended and double-ended converters. These converters may or may not have a high-frequency transformer for isolation.

Nonisolated Single-Ended PWM Converters.

The basic nonisolated single-ended converters are (a) buck (step-down), (b) boost (step-up), (c) buck-boost (step-up or step-down, also referred to as flyback), and (d) 'Cuk converters (Figure 6. 5. 2). The first three of these converters have been discussed in Section 30. 2. The 'Cuk converter provides the advantage of

nonpulsating input-output current ripple requiring smaller size external filters. Output voltage expression is the same as the buck-boost converter and can be less than or greater than the input voltage. There are many variations of the above basic nonisolated converters, and most of them use a high-frequency transformer for ohmic isolation between the input and the output. Some of them are discussed below.

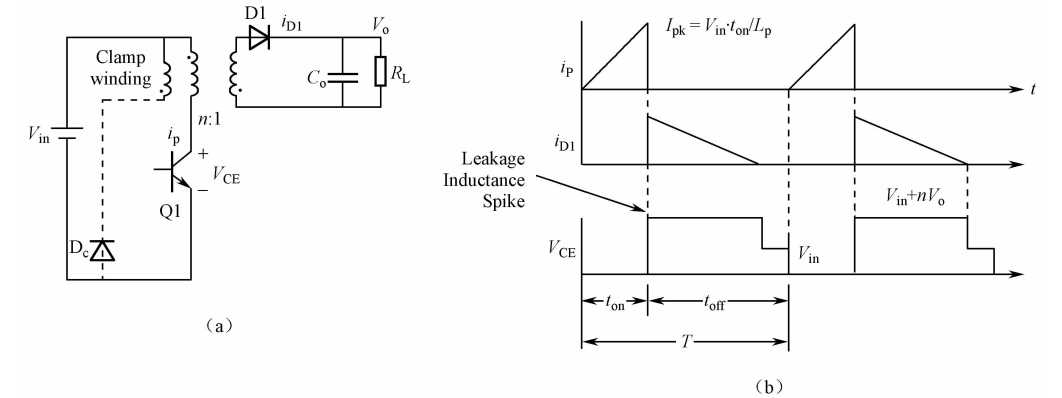


Figure 6. 5. 3 (a) Flyback converter. The clamp winding shown is optional and is used to clamp the transistor voltage stress to $V_{in} + nV_o$. (b) Flyback converter waveforms without the clamp winding. The leakage inductance spikes vanish with the clamp winding.

Isolated Single-Ended Topologies

1. *The flyback converter* (Figure 6. 5. 3) is an isolated version of the buck-boost converter. In this converter, when the transistor is on, energy is stored in the coupled inductor (not a transformer), and this energy is transferred to the load when the switch is off.

Some of the advantages of this converter are that the leakage inductance is in series

with the output diode when current is delivered to the output, and, therefore, no filter inductor is required; cross regulation for multiple output converters is good; it is ideally suited for high-voltage output applications; and it has the lowest cost.

Some of the disadvantages are that large output filter capacitors are required to smooth the pulsating output current; inductor size is large since air gaps are to be provided; and due to stability reasons, flyback converters are usually operated in the DCM, which results in increased losses. To avoid the stability problem, flyback converters are operated with current mode control explained earlier. Flyback converters are used in the power range of 20 to 200 W.

2. *The forward converter* (Figure 6.5.4) is based on the buck converter. It is usually operated in the CCM to reduce the peak currents and does not have the stability problem of the flyback converter. The HF transformer transfers energy directly to the output with very small stored energy. The output capacitor size and peak current rating are smaller than they are for the flyback. Reset winding is required to remove the stored energy in the transformer. Maximum duty cycle is about 0.45 and limits the control range. This topology is used for power levels up to about 1 kW.

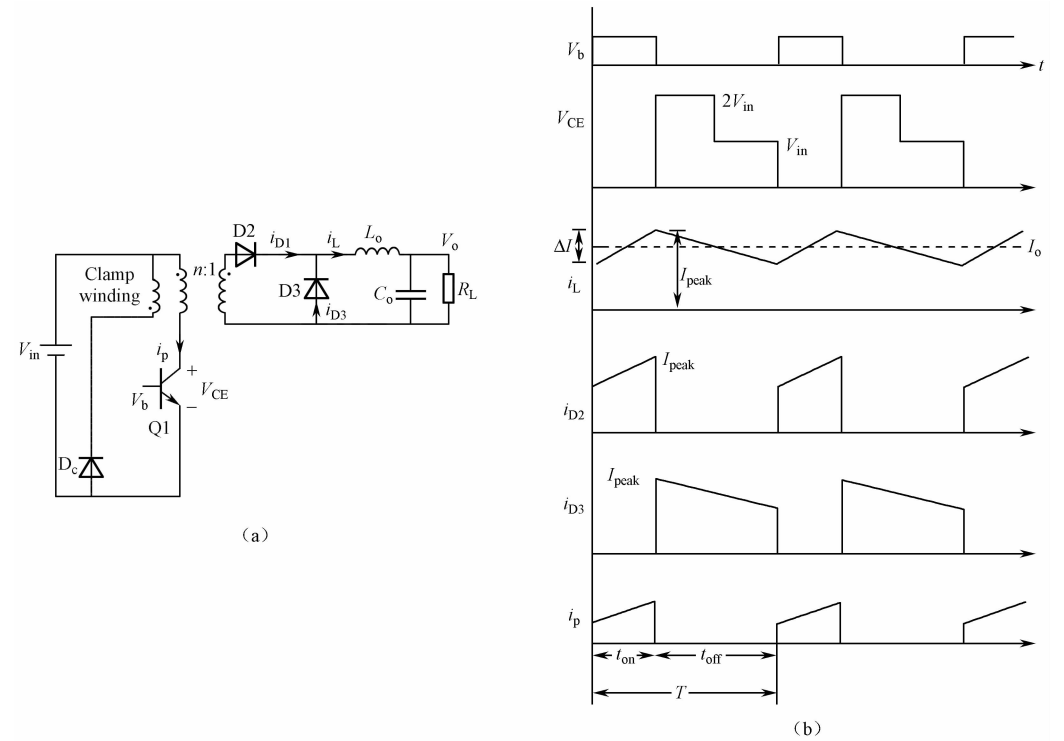


Figure 6.5.4 (a) Forward converter. The clamp winding shown is required for operation.
(b) Forward converter waveforms.

6.5.2 Specialized English Words

aerospace 航天	open-loop 开环
harmonic distortion 谐波畸变	line regulation 电源调整率
power factor 功率因数	clamp 箝位
line frequency 工频	leakage inductance spike 漏感电压尖峰
ripple 纹波	compensation networks 补偿网络
resonant converter 谐振变换器	single-ended converters 单端变换器
electromagnetic interference (EMI) 电磁干扰	double-ended converters 双端变换器
in nature 实际上,本质上	buck(step-down) converter 降压变 换器
duty cycle 占空比	boost(step-up) converter 升压变换器
ramp 锯齿波	cuk converter 丘克变换器
feedforward 前馈	flyback converter 反激式变换器
magnetics 磁元件	forward converter 正激式变换器
audio susceptibility 音频敏感性	leakage inductance 漏感
feedback 反馈	filter inductor 滤波电感
hold-up time 保持时间	in proportion to 与……成比例

6.5.3 Notes

[1] Also, the switching stresses are high with the generation of large electromagnetic interference (EMI), which is difficult to filter. 句中的“which is difficult to filter”为非限定性定语从句,说明 EMI。“large”在这里译为“大量的”比较合适。全句译为“还有,随着难以消除的大量的电磁干扰(EMI)的产生,开关应力也会很高。”

[2] In this method, a second inner control loop compares the peak inductor current with the control voltage which provides improved open-loop line regulation [Figure 6.5.1(c)]. 句中“compares...with...”意为“将……与……比较”,“which provides improved...”为定语从句,修饰“the control voltage”,其中“improved”在这里是过去分词做定语。全句译为“在这种控制方式中,第二个内部控制环将电感的峰值电流和用来改善开环电源调整率的控制电压做比较[见图 6.5.1(c)]。”

[3] An additional advantage of this method is that the two-pole second-order filter is reduced to a single-pole (the filter capacitor) first-order filter, resulting in simpler compensation networks. 句中“that the two-pole second-order filter...”为表语从句,“resulting in simpler compensation networks”为现在分词做结果状语。全句译为“这种方式的另外一个优点就是用单级(滤波电容)一阶滤波器替代两极二阶滤波器,这样补偿网络就更简单了。”

6.5.4 Reference Translation

直流电源

在许多工业和航空应用以及消费产品中都要用到电源。有些要求电源体积小、重量轻、成本低以及电能转换效率高。除此之外,有些电源还要求电源和负载要有电气隔离、输入输出波形的谐波畸变要低,以及若电源为交流电,则要求功率因数要高。某些特别的电源还要求控制电能的方向。

需求的电源基本上有两种类型:直流电源和交流电源。直流电源的输出为稳定的或者是可控的直流电,而交流电源的输出为交流电。这些电源的输入为交流电或者直流电。

如果采用的是交流电源,那么可采用已在 30.2 节中解释过的交流到直流的变换器。在这些变换器中,只能采用大体积的工频变压器来实现电气隔离。交流电源经二极管整流器而得到一个不可控的直流电源,然后采用直流变换器得到可控的直流电源。直流变换器中可采用高频(HF)变压器来实现输入电源和负载之间的电气隔离。这种高频变压器相对工频变压器来说,体积小、重量轻且成本低。无论输入电源是直流(如电池)或者是交流,直流变换器都是直流电源中的一个重要组成部分,在本节里将对它进行分析。

直流电源可以泛分为线性电源和开关电源。

线性电源是最老的和最简单的一种电源类型。它是通过将多余的电压降落到串联晶体管(也称为串联调节器)上来实现输出电压的调节的。这种电源的输出纹波非常小,理论上为零噪声,保持时间长(典型的为 $1\sim 2\text{ ms}$),响应快速。线性电源有以下缺陷:效率非常低、只能在 60 Hz 的交流侧进行电气隔离,体积大且重,而且一般来说仅有一个输出。不过它们在非常小的稳压电源和某些特殊的应用(如磁电源)中仍然有使用。三端线性稳压集成电路(ICs)很多(如 $\mu\text{A}7815$ 输出为 $+15\text{ V}$, 1 A),且易于使用,还有内置负载短路保护。

采用工作于开关状态的半导体器件的开关电源效率高、体积小且重量轻。由于有高速开关器件、高频磁元件和高频电容以及高速控制 IC,开关电源得到了广泛应用。它们可以进一步分为脉宽调制(PWM)变换器和谐振变换器,下面将给以阐述。

脉宽调制变换器

这类变换器采用方波脉宽调制来实现电压的调整。通过调节电力半导体开关的占空比来调整平均输出电压。实际通过开关器件和输出的电压波形为方波,这样当开关频率增加时高频的开关损耗将会增加。还有,随着难以消除的大量的电磁干扰(EMI)的产生,开关应力也会很高。不过这种变换器易于控制,易于理解且有很宽的负载控制范围。

下面我们将讨论 PWM 变换器的控制方式。

控制方式。PWM 变换器一般频率固定,占空比可调。根据占空比的情况,变换器可以工作在电流连续模式(CCM)或者电流断续模式(DCM)。如果流过输出电感的电流总不为零,则变换器工作在 CCM,否则工作在 DCM。

下面简述三种控制方式。

1. 直接占空比控制。这是最简单的控制方式。用固定频率的锯齿波和控制电压[见

图 6.5.1(a)]比较以获得一个可变的占空比来作为晶体管基极驱动信号。这是最简单的控制方法。这种控制方法的缺点在于:(a)没有电压前馈,这样无法预估输入电压变化产生的影响,对于输入变化突变情况反应迟钝,声频敏感度低,必须采取高增益的方法才能达到指标要求;(b)动态响应差。

2. 电压反馈控制。这种方法中锯齿波的幅值变化直接和输入电压成正比[见图 6.5.1(b)]。开环调节非常好,解决了 1(a)中的问题。

3. 电流模式控制。在这种控制方式中,第二个内部控制环将电感的峰值电流和用来改善开环电源调整率的控制电压做比较[见图 6.5.1(c)]。这种方式解决了方式 1 直接控制占空比方式的所有问题。这种方式的另外一个优点就是用单级(滤波电容)一阶滤波器替代了两极二阶滤波器,这样补偿网络就更简单了。

上述三种控制方式在下面将要分析的所有的 PWM 变换器中都有应用。

PWM 变换器可分为单端变换器和双端变换器。它们或许有用于隔离的高频变压器,或许没有。

非隔离型单端 PWM 变换器

基本的非隔离型的变换器有(a)降压,(b)升压,(c)升降压,(d)丘克变换器[见图 6.5.2]。前面三种变换器在 30.2 节已经讨论过了。丘克变换器的输入输出电流纹波变化小且外部滤波器尺寸更小。其输出电压的表达式和升降压变换器的一样,可以大于或者小于输入电压。上述非隔离型的变换器的种类可以有很多种变化,大多数采用高频变换器来对输入和输出进行电阻隔离。下面我们将讨论其中的几种。

单端隔离型变换器

1. 反激式变换器是带隔离的升降压变换器。在该变换器中[见(图 6.5.3)],当晶体管导通时,电能存储在耦合电感(而不是变压器)中,当晶体管断开时,电能传递给负载。

该变换器的优点是,当电流流向负载时漏感和输出二极管串联,因此无需滤波电感;采用交叉调整的多输出变换器的效果很好,非常适合高压输出的应用,且成本最低。

该变换器的缺点是,为了使脉动的输出电流平滑,需要大的输出滤波电容;由于稳定性的原因,反激式变换器通常工作在电流断续状态,这样会导致损耗增加(这在前面已经解释过)。反激式变换器常用功率范围在 20~200 W 之间。

2. 正激式变换器[见(图 6.5.4)]基于降压式变换器。为减少峰值电流,它通常工作在电流连续状态且没有反激式变换器那种稳定性方面的问题。高频变压器几乎不存储电能而直接将能量输出。其输出电容的尺寸和峰值电流额定值都小于正激式的。如果要清除变换器内存储的电能,需要将绕组重置。最大占空比大约为 0.45,因此控制范围有限。这种结构的电路功率最大可达 1 kW。

6.5.5 Reading Materials

Overview of Inverter(逆变器)

Inverters are used to create single or polyphase(多相的) AC voltages from a DC

supply. In the class of polyphase inverters, three-phase inverters are by far the largest group. A very large number of inverters are used for adjustable speed motor drives. The typical inverter for this application is a “hard-switched” voltage source inverter producing pulse-width modulated (PWM) signals with a sinusoidal fundamental. Recently research has shown detrimental effects on the windings and the bearings resulting from unfiltered PWM waveforms and recommend the use of filters. A very common application for single-phase inverters are so-called “uninterruptable power supplies” (UPS) for computers and other critical loads. Here, the output waveforms range from square wave to almost ideal sinusoids. UPS designs are classified as either “off-line” or “online”. An off-line UPS will connect the load to the utility for most of the time and quickly switch over to the inverter if the utility fails. An online UPS will always feed the load from the inverter and switch the supply of the DC bus instead. Since the DC bus is heavily buffered with capacitors, the load sees virtually no disturbance if the power fails.

In addition to the very common hard-switched inverters, active research is being conducted on softswitching (软开关) techniques. Hard-switched inverters use controllable power semiconductors to connect an output terminal to a stable DC bus. On the other hand, soft switching inverters have an oscillating intermediate circuit and attempt to open and close the power switches under zero-voltage and or zero-current conditions.

A separate class of inverters are the line commutated inverters for multimewatt power ratings, that use thyristors (also called silicon controlled rectifiers, SCRs). SCRs can only be turned “on” on command. After being turned on, the current in the device must approach zero in order to turn the device off. All other inverters are self-commutated, meaning that the power control devices can be turned on and off. Line commutated inverters need the presence of a stable utility voltage to function. They are used for DC-links between utilities, ultralong distance energy transport, and very large motor drives. However, the latter application is more and more taken over by modern hard-switched inverters including multilevel inverters.

Modern inverters use insulated gate bipolar transistors (IGBTs) as the main power control devices. Besides IGBTs, power MOSFETs are also used especially for lower voltages, power ratings, and applications that require high efficiency and high switching frequency. In recent years, IGBTs, MOSFETs, and their control and protection circuitry have made remarkable progress. IGBTs are now available with voltage ratings of up to 3300 V and current ratings up to 1200 A. MOSFETs have achieved on-state resistances approaching a few milliohms. In addition to the devices, manufacturers today offer customized control circuitry that provides for electrical isolation, proper

operation of the devices under normal operating conditions and protection from a variety of fault conditions. In addition, the industry provides good support for specialized passive devices such as capacitors and mechanical components such as low inductance bus-bar assemblies to facilitate the design of reliable inverters. In addition to the aforementioned inverters, a large number of special topologies are used. A good overview is given by Gottlieb .

Fundamental Issues

Inverters fall in the class of power electronics circuits. The most widely accepted definition of a power electronics circuit is that the circuit is actually processing electric energy rather than information. The actual power level is not very important for the classification of a circuit as a power electronics circuit. One of the most important performance considerations of power electronics circuits, like inverters, is their energy conversion efficiency. The most important reason for demanding high efficiency is the problem of removing large amounts of heat from the power devices. Of course, the judicious use of energy is also paramount(极为重要的), especially if the inverter is fed from batteries such as in electric cars. For these reasons, inverters operate the power devices, which control the flow of energy, as switches. In the ideal case of a switching event, there would be no power loss in the switch since either the current in the switch is zero (switch open) or the voltage across the switch is zero (switch closed) and the power loss is computed as the product of both. In reality, there are two mechanisms that do create some losses, however; these are on-state losses and switching losses. On-state losses are due to the fact that the voltage across the switch in the on state is not zero, but typically in the range of 1 to 2 V for IGBTs. For power MOSFETs, the on-state voltage is often in the same range, but it can be substantially below 0.5 V due to the fact that these devices have a purely resistive conduction channel and no fixed minimum saturation voltage like bipolar junction devices (IGBTs). The switching losses are the second major loss mechanism and are due to the fact that, during the turn-on and turn-off transition, current is flowing while voltage is present across the device. In order to minimize the switching losses, the individual transitions have to be rapid (tens to hundreds of nanoseconds) and the maximum switching frequency needs to be carefully considered.

In order to avoid audible noise being radiated from motor windings or transformers, most modern inverters operate at switching frequencies substantially above 10 kHz.

Part 7 Miscellaneous

7.1 What is Electrical Engineering?^[1]

7.1.1 Text

Electrical Engineering, affectionately known as “EE” or “Double E”, is an exciting and fascinating field of study which encompasses almost all walks of daily life in this “Hi-Tech” era. The alarm clock which wakes you up with a soft warble, the coffee maker which almost hands you your early morning potion as you walk into the kitchen, the almost-magical wireless garage door opener, the cassette/radio in your car that plays soothing music to keep you calm as you work your way through rush-hour traffic, the elevator that takes you up to your office on the umpteenth floor at the touch of a button, the ubiquitous telephone you cannot imagine living without, the computer on your desk that does all your arithmetic and more, and ... the list is endless ... All of these are EE products, with decades of dedicated efforts of electrical engineers behind them^[2]. You find EE concepts and tools being used in radio, television, telecommunication, satellite and space technology, monitoring and control of various equipment and processes, computers information processing, medical imaging and diagnosis, electronics and instrumentation, data recording and processing, and just about any hi-tech tool or gadget you use.

How did all of these come to be? How does electric power get transmitted along those lines? How does a TV put up all those beautiful, moving pictures? How does a cellular phone work? How does a computer add or multiply numbers so fast? How does a computed tomography scanner show your innards?

The Department of Electrical and Computer Engineering at The University of Calgary (U of C) offers a comprehensive, accredited, undergraduate B. Sc. EE program (with an optional Minor in Computer Engineering), as well as advanced, high-quality M. Sc. , M. Eng. , and Ph. D. programs. The Department was established in 1966, and today has a 24 academic staff members, 18 technical and research staff, about 85 students in each of its 3rd and 4th year classes, and about 70 graduate students. With an annual operating budget of about \$2,700,000, research funds totaling about \$2,300,000 annually, and an equipment inventory running to almost \$4,000,000, we are now of the leading engineering departments in Canada^[3]. Many of our professors are world-renowned, and have been recognized by awards for innovation and fellowships of learned societies.

The Department houses the Electrical Engineering Students' Society (EESS, popularly known as "The Zoo"), the IEEE (Institute of Electrical and Electronics Engineers, the international EE professional organization) Student Branch, and the IEEE McNaughton Student Learning Centre. These societies provide opportunities for all-round development of our students.

What will you learn if you join us? Well, the first one-and-a-half years of engineering studies are common for all branches, including the basics of physics, mathematics, chemistry, computers electric circuits and systems, strength of materials, etc.

The real B. Sc. EE program takes off in the second term of the second year, with introductory courses in the areas of circuit analysis, signal and systems, digital circuit design, data structures and software development.

The 3rd year EE program introduces basic concepts of electric power, circuits, systems, and machines; electronic devices and design of integrated circuits and networks; electromagnetic fields and radiation; communication of radio, television, and computer signals; control systems; digital system design, assembly language programming, and interfacing of the development of microcomputer-based control systems; and computer data structures.

The fourth year program gets more interesting, exciting, and challenging! Fourth year EE students learn about digital integrated electronics, instrumentation, digital control, power systems, power electronics, digital signal processing, and advanced communication topics such as digital communication, microwaves, and fiber-optics. Students in the Computer Engineering Minor program get to know more about computer systems, computer architecture, digital filters and signal processing, computer graphics and visualization, computer communication, and computer operating systems. We offer a wide range of electives so that students may choose courses which tickle their intellect. For those who are more adventurous, we offer a project course where they may work on an advanced-level project on a one-on-one basis with a professor!

Graduate Studies and Research. Did you say all above is not enough? Right on! We have more! Knowledge is such that you learn a little, and you develop a thirst and yearn for more! If one's brain itches for more, there are certainly the right kinds of problems to brood over in EE graduate studies. Our professors and research students work on such problems as modeling and control of large-scale power generation and distribution systems; design, analysis, and optimization of communication networks including microwave, fiber-optic, satellite, and computer communication systems; optimization of engineering systems; design of complex VLSI (very large scale integration) circuits for computers and signal processing; electronic instrumentation for various applications including oil exploration; computer graphics; biomedical imaging, image processing,

and signal analysis, with applications in diagnosis of knee joint problems, breast cancer, etc; design of computer hardware and software systems; and just about anything else! We have a first-class group of staff members with state-of-the-art equipment in our Department to find answers to the most challenging problems in this hi-tech age.

Engineering-A career to look forward to! According to a 1993 salary survey conducted by the Association of Professional Engineers, Geologists, and Geophysicists of Alberta (APEGGA), engineers in Alberta earn an average starting salary of about \$36,000 per annum. Engineers are a competitive lot, and many rise to senior management and specialist levels, where the average annual salary is more than \$100,000!

Interested? Many scholarships are available for engineering studies at The U of C, such as the APEGGA Alexander Clayton Milroy Scholarships, the Louise McKinney (Alberta Heritage) Scholarships, Suncor Inc. Scholarship, UMA Group Scholarship, and the Bill Howard Memorial Foundation Awards.

Graduate students in our Department are offered financial support of about \$14,000 per annum. Many scholarships are also available; 57 scholarships were secured by our graduate students in 1993 from the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Microelectronics Centre (AMC), Alberta Government Telephones (AGT), Canadian International Development Agency (CIDA), TR Labs, and other agencies.

An interesting lot! Aren't we? So, the next time you change your light bulb... well, let us be more up-to-date... the next time you pick up your phone, or play your cassette/radio/TV, or flick your computer/calculator on, or ride the C-Train or an elevator, or hear about an advanced medical imaging device, think of ELECTRICAL ENGINEERING!

7.1.2 Specialized English Words

affectionately 亲昵地

fascinating 迷人的

B. Sc. 理(科)学士学位(Bachelor of Science 之缩写)

M. Sc. 理科硕士学位(Master of Science 之缩写)

M. Eng. 工科硕士学位(Master of engineering 之缩写)

Ph. D. 博士学位(Doctor of Philosophy 之缩写)

per annum 每年

geologist 地质学家

geophysicist 地球物理学家

encompass 包含

walks 方法,路径,方面

warble (鸟的)柔和叫声

potion 饮料(旧字),有魔力的饮料

garage 车库

soothe 平静,镇定,安慰

soothing music 令人舒畅的音乐

umpteen 数不清的,无数的
 ubiquitous 无处不在的
 gadget 小机械,小器具
 innards 肠胃(口语)
 cellular phone 蜂窝电话,手机
 computed tomography 计算机断层扫描(图像)(CT)
 comprehensive 全面的,综合的
 comprehension 理解(力)
 accredited 公认的,官方认可的
 minor (in) 副修课程
 minority 少数(民族)
 budget 预算,专用开支
 inventory 清单,目录
 award 奖金,助学金;颁发,授予
 renowned 著名的
 recognize 承认,授予荣誉,奖励
 innovation 革新,新技术,新思想
 fellowship 会员资格,研究生奖学金,
 友情
 learned societies 学术团体

house 给……住所
 take off 起飞,突然离去,突然大受欢迎
 elective 选修课;选修的,有选举权的
 tickle 骚痒,激发,使……有兴趣
 intellect 智力
 adventurous 喜欢冒险的,喜欢探索的
 adventure 奇遇,冒险
 thirst for 渴望,渴求
 yearn for 渴望,渴求
 itch 渴望,渴求,痒
 brood 孵雏
 conduct 指导,主持
 survey 观察,综述,调查
 a competitive lot 一群求胜心切的人,
 一群上进心强的人
 secure 获得,给予安全
 flick…with… 用……轻打……
 heritage 文化遗产
 inherit 继承
 inheritance 继承,遗物,遗产

7.1.3 Notes

[1]本文是一篇向社会,主要是中学毕业生,介绍大学相关专业的招生宣传材料。它和正规专业介绍及招生简章不同。内容是客观真实的,文字上具有一些文学色彩,口语的运用和对话口吻的行文方式显得亲切自然,颇有鼓动煽情味道,但又不是广告。

[2]The alarm clock which wakes you up with a soft warble, the coffee maker which almost hands you your early morning potion as you walk into the kitchen, the almost-magical wireless garage door opener, the cassette/radio in your car that plays soothing music to keep you calm as you work your way through rush-hour traffic, the elevator that takes you up to your office on the umpteenth floor at the touch of a button, the ubiquitous telephone you cannot imagine living without, the computer on your desk that does all your arithmetic and more, and … the list is endless … All of these are EE products, with decades of dedicated efforts of electrical engineers behind them. 在这个长句中,前面是一串罗列的名词,其中一些带有后置定语从句。而“the ubiquitous telephone”的后置定语从句“you can’t imagine living without”中现在分词短语“living without”是“imagine”的状语。换成如下解释性说法意思就更清楚了:“You cannot

imagine life without telephone.”全句可译为“清晨用柔和鸟鸣声把你唤醒的闹钟；是把香浓美味的咖啡直接递到刚在进厨房的你的手中的咖啡壶；就像变魔术一样的无线电车库开门器；当你在上下班交通拥堵不堪的路上艰难开车前行的时候，在车上播放出优雅的音乐，让你烦躁的心情平静下来的录放/收音机；只要一按电钮，就把你送达位于高高楼上的办公室的电梯；一旦生活中没有了它就无法想象的无处不在的电话；你桌上的那台包揽了你的各种计算以及别的事情的计算机，等等……所有这些都是 EE 产品，是它背后的电气工程师们几十年来不懈努力的结晶。”

[3] With an annual operating budget of about \$ 2,700,000, research funds totaling about \$ 2, 3000, 000 annually, and an equipment inventory running to almost \$ 4,000,000, we are now of the leading engineering departments in Canada. 句中“with …to almost \$ 4,000,000”系介词短语，用做状语。注意句中有多个现在分词的运用，均为定语，“operating”，“leading”，“engineering”为前置定语，放在所修饰词的前面；“totaling”，“running”为后置定语，放在所修饰词的后面。全句可译为“系的年度运行预算约270 000加元，年度科研经费约 2 300 000 加元，仪器设备费用达到近 4 000 000 加元。我系现在是全加拿大领先的大学工程系部之一。”

7.1.4 Reference Translation

什么是电气工程

电气工程，被亲切地称为“EE”或者“双E”，是一个令人着迷、令人激动的学习领域。在当今这个高科技时代里，电气工程几乎涉及日常生活的所有方面。清晨用柔和鸟鸣声把你唤醒的闹钟；把香浓美味的咖啡直接递到刚跨进厨房的你的手中的咖啡壶；像变魔术一样的无线电车库开门器；当你在上下班交通拥堵不堪的路上艰难开车前行的时候，播放出优雅的音乐，让你烦躁的心情平静下来的车载录放/收音机；只要一按电钮，就把你送达位于高高楼上的办公室的电梯；一旦生活中没有了它就无法想象的无处不在的电话；你桌上的那台包揽了你的各种计算以及别的事情的计算机……所有这些都是 EE 产品，是它背后的电气工程师们几十年来不懈努力的结晶。你可以看到 EE 的理念和产品被用于无线电、电视、通信、卫星和空间技术、各种装置和生产过程的监视与控制、计算机、信息处理、医学成像和诊断、电子器件与仪表、数据记录与处理，真可谓是从高科技产品到你手边的小器具，无所不包。

这一切是如何形成的？电能是如何沿着电缆传送的？电视机是如何展现出那些漂亮的活动画面的？手机是怎样工作的？为什么计算机运算如此神速？计算机断层扫描仪(CT)是怎样把人的内脏显现出来的？

卡尔加里大学电气与计算机工程系提供综合性的正规本科工学学士教学课程(还有计算机工程辅修课程可选)以及先进优质的理科和工科硕士及博士学位课程。本系成立于1966年，现有24名专业教师，18名技术和研究人员。三年级和四年级各有学生约85名，研究生约70名。系的年度运行预算约270 000加元，年度科研经费约2 300 000加

元。仪器设备费用达到近 4 000 000 加元。我系现在是全加拿大领先的大学工程系部之一。我系拥有很多世界著名教授,他们因创新成果而获奖并成为各学术团体成员。

我系有电气工程学生会(EESS,俗称“动物园”)、IEEE(电气与电子工程师学会,国际电气工程专业组织)学生分会以及 IEEE 麦克劳顿学生学习中心。这些学会组织为学生的全面发展提供了多种机会。

到我们系能学些什么? 所有专业方向的学生,前一年半(学年)的课程是一样的,包括物理、数学、化学、计算机、电路与系统、材料力学等基础课。

工程学士的真正课程是从第二(学)年的下学期的导论性课程开始的,涉及电路分析、信号与系统、数字电路设计、数据结构及软件开发。

EE 三年级的课程讲述的是电能、电路、电网系统、电机的基本理论;电子器件、集成电路与网络设计、电磁场与辐射;无线电、电视与计算机信号通信;控制系统;微机控制系统的开发、数字系统设计、汇编语言编程及接口;计算机数据结构。

第四年的课程更是饶有趣味,更令人激动且更具挑战性!四年级学生的学习涉及数字集成电子学、仪表、数字控制、电力系统、电力电子学、数字信号处理以及像数字通信、微波和光纤这样的高级通信论题。辅修计算机工程的学生要学计算机系统、计算机组成、数字滤波与信号处理、计算机图形与可视化、计算机通信及计算机操作系统。还开有大量选修课供学生选择以激发他们的聪明才智。对于探索欲望强烈的学生,我们可以提供专门的项目课程,由教授一对一地指导他们钻研高水平课题。

研究生的学习和研究。你是不是认为上面所有这些还不够?接着瞧吧!还有更多的呢!所谓知识,就是当你学了一点,你就渴望得到更多。如果哪位还盼望更多的东西,我们的研究生学习一定有适合他的课题供他探讨。我们系的教授和研究生研究的是这样的一些问题:大规模发电配电系统的建模与控制;包括微波、光纤、人造卫星和计算机通信系统在内的通信网络的设计、分析和优化;工程系统的优化;用于计算机和信号处理的复杂 VLSI(极大规模集成)电路的设计;包括石油勘探在内的各种应用的电子仪表,计算机影像学;用于膝关节疾病、乳腺癌等的诊断的生物医学成像、图像处理、信号分析;计算机硬件软件设计。可以说无所不包!我们拥有一流的师资和一流的设备来探寻这个高科技时代中最具挑战性问题的答案。

工程学科——令人企盼的人生!根据由阿尔伯塔专业工程师、地质学家和地球物理学家学会(APEGGA)主导的一项 1993 年工资调查表明,阿尔伯塔省的工程师的年均起薪为 36 000 加元。工程师是一个富于竞争力的群体,很多人升迁到高级管理层或专家层,平均年薪在 100 000 加元以上!

感兴趣了吗?卡尔加里大学有很多向工科学生提供的奖学金。例如 APEGGA 的亚历山大·克雷顿·米尔诺奖学金、路易丝·麦克米伦(阿尔伯塔遗产)奖学金、圣可公司奖学金、UMA 集团奖学金以及比尔·赫伍德纪念基金奖学金等。

电气系的研究生每年可以获得大约 14 000 加元的官方经费资助。同样也有很多种奖学金:1993 年我系研究生共获得 75 份奖学金,来自加拿大国家科学和工程研究会(NSERC)、阿尔伯塔微电子中心、阿尔伯塔国营电话公司、加拿大国际开发署、TR 实验室

及其他机构。

令人感兴趣的一群！难道不是吗？所以下次当你更换灯泡时，唔，让我们更现代一点——下次当你拿起电话的时候，当你打开录放/收音机/电视机的时候，当你轻快地敲击计算机/计算器键盘的时候，当你乘坐轻轨或电梯的时候，当你听说到高级医学成像设备的时候，就好好想想“电气工程”吧！

7.2 The Belief of HIMA^[1]

7.2.1 Text

For more than 40 of the past 100 years of our company history we have been involved exclusively with the safety of critical processes. As a pioneer of safety systems and as an independent supplier of safety technology, we turn our customers' requirements into tailored solutions with a guaranteed future. Safety, availability, efficiency and maximum flexibility are the cornerstones that determine the way we act.

Solutions are our greatest strength. Whether process, machine or building safety, we offer our customers complete solutions for safety-related applications, all from a single source. Because our staff are so highly qualified and experienced, we are better placed than any other company to advise our customers on safety concepts and matters of functional safety.

One of our decisive competitive advantages is our independence. We reinvest an above-average amount in the ongoing development of our core competencies, enabling us to offer our customers solutions of the very highest caliber. HIMA supports a department at the University of Kassel in Germany which deals with the safety philosophies of the future.

At HIMA, Independent Open Integration is not just a hollow term, it's a belief. We have employed innovative technology to create the ideal link to every common DCS system, and we consider ourselves an integral part of these systems. HIMA has a team dedicated solely to this discipline. And furthermore, our long-term, stable corporate culture means our customers can rest assured that, by choosing HIMA as a safety partner, they have made the best technical and commercial decision for their plant and operations that will be supported well into the future^[2].

HIMA's excellent long-term outlook is a direct result of the exceptional quality of our development, engineering, products and services. As an independent, autonomous supplier of safe automation solutions we will continue to grow profitably and provide unparalleled service to our customers both in Germany and abroad. From safety consulting all the way to world-class safety solutions, we offer services that cement our position as Number One in the safety technology market.

The longevity of HIMA is marked by our 100th birthday milestone and highlights our success in meeting the demands of today’s rapidly evolving industrial automation and safety marketplace.

7.2.2 Specialized English Words

tailored 订制的	hollow 空洞的
cornerstone 基石	rest assured that 放心
University of Kassel 卡赛尔综合大学	unparalleled 无双的
competencies 竞争力	cement 加固
caliber 标准	longevity 长寿

7.2.3 Notes

- [1]这是一家公司的自我宣传介绍材料的一部分。这种材料都是正面介绍宣传自己的长处和优势。行文庄重,注意修辞。虽不乏溢美之词,但其基本内容必须符合事实。
- [2]And furthermore, our long-term, stable corporate culture means our customers can rest assured that, by choosing HIMA as a safety partner, they have made the best technical and commercial decision for their plant and operations that will be supported well into the future. 这是一个由三个句子组成的复合句。“our long-term, stable corporate culture”为主句主语,“means”为谓语动词,“our customers can rest assured”为其宾语从句。“assured that, by choosing HIMA as a safety partner, they have made the best technical and commercial decision for their plant and operations”则是从句的宾语从句,而“that will be supported well into the future”又是“plant and operations”的定语从句。全句可译为“此外,我们悠久稳定的企业文化意味着我们的客户一旦选择了 HIMA 作为安全伙伴,就可以安心释怀,因为他们为自己的工厂和工厂的运作做出了最好的技术和商业选择,可以得到可靠的支持,确保自己的未来。”

7.2.4 Reference Translation

HIMA 的信仰

在 HIMA 公司过去一百多年的历史中,我们在重大生产过程安全领域独领风骚已超过 40 年。作为安全系统的先驱以及安全技术的独立供应商,我们按客户的需求为他们量身订制解决方案,确保其前程无忧。安全性、可用性、有效性及最大的灵活性是我们赖以行动的基石。

解决方案是我们的力量所在。无论是过程、设备还是建筑的安全,我们都能根据客户的安全需求提出完整的解决方案。所有方案均有一共同的源头,即源于我们员工的高素质 and 宝贵的经验。因而在向客户提供安全理念和安全产品的时候,我们都处在比任何其他公司更强有力的位置上。

我们决定性的竞争优势之一在于我们的独立性。我们对开拓核心竞争力的新项目进行了大量再投资,使我们得以按照最高标准为客户提供解决方案。HIMA 资助德国卡塞尔大学的一个系,进行有关未来安全理念的研究。

在 HIMA, **独立—开放—集成**不是空洞的名词而是一种信仰。我们用革命性的技术去创新每一个普普通通的 DCS 系统,并把自己视为这些系统中的不可分的一部分。HIMA 的团队全身心地投入到这一信念之中。此外,我们悠久稳定的企业文化意味着我们的客户一旦选择了 HIMA 作为安全伙伴,就可以安心释怀,因为他们为自己的工厂和工厂的运作做出了最好的技术和商业选择,可以得到可靠的支持,确保自己的未来。

HIMA 辉煌的长远发展前景是我们开发、施工、产品和服务的优异品质的必然结果。作为一家独立的自动化安全方案的供应商,我们将继续良性发展,为德国及全球客户提供独一无二的服务。从安全咨询到世界一流安全方案,我们提供的各种服务使我们稳居安全技术市场首屈一指的宝座。

HIMA 的百年寿辰标志着它的长寿,凸现了我们在满足当今迅速发展的工业自动化和安全市场需求方面的成功。

7.2.5 Reading Materials

HIMA (Shanghai) Safety System Co. , Ltd

HIMA (Shanghai) Safety System Co. , Ltd(上海黑马安全自动化系统有限公司) was founded in 2002 as a subsidiary(分公司) of HIMA Paul Hildebrandt GmbH + Co KG. It is responsible for all business in china including product sales, engineering design, technical training, system integration, maintenance service and spare parts(备品,备件) supply. It locates in Waigaoqiao Free Trade Zone Shanghai Pudong District (上海浦东外高桥保税区), and owns more than 6,000 square meters for office and project center. HIMA Shanghai sets up four offices in Beijing, Shenyang, Shenzhen and Chengdu, develops a powerful sales and technical service network.

HIMA Company was founded in 1908 in Germany; it is a world-famous and professional manufacturer for safety-related control system. Since the first set of TÜV certificated(经过认证的) fail-safe interlock system Planar(故障安全型控制系统 Planar) in the world launched in 1970, HIMA has always been at the cutting-edge in the field of safety-related systems. She has led the technology development of safety-related control systems for three generations in the world.

HIMA has been being devoted to the development and application of safety-related control systems and always keep a leading technical position in this field. In the past over 30 years, There were more than 20,000 sets of safety-related systems produced by HIMA globally used in many business and fields, including chemical and petrochemical industry, offshore oil platform(海上石油平台), long-distance oil-gas pipeline(油气长输

管线), metallurgy (冶金), power generation, machinery, transportation and large public building.

7.3 A Letter about Scientific Communications^[1]

7.3.1 Text

THE UNIVERSITY OF CALGARY
FACULTY OF ENGINEERING

Department of Electrical and Computer Engineering

1994-12-12

Professor XXX

Department of Mechano-electronic Engineering

Wuhan Food Industry College

Wuhan, Hubei 430022

People's Republic of China^[2]

Dear Professor XXX:

Please refer to your letter of 1994-11-20 enquiring about the possibility of coming to the University of Calgary as a visiting scholar for a period of six months and doing research in collaboration with our research group. You state that all your travelling expenses and living costs will be borne by your government.

As you are probably aware, in my research group we are doing research in the area of adaptive control, fuzzy logic control and neural network based control for application to power systems. If you are interested in doing research in this area, I will be glad to accept you as a visiting scholar for a period of six months.

You should be aware that the cost of living in Canada is quite high and you should make sure that you have enough funds. We also advise all our foreign visitors to get health insurance during their stay here.

I can accept you in our laboratory at anytime as you will be joining our on-going research activity. If you are interested in coming to Calgary, please contact the Canadian Embassy in Beijing for a suitable visa.

Yours sincerely,
XXX, Professor
Electrical & Computer
Engineering Department^[3]

7.3.2 Specialized English Words

visiting scholar	访问学者	health insurances	健康保险,医疗保险
living costs	生活费	embassy	大使馆
power system	电力(供应)系统,能源系统	visa	签证

7.3.3 Notes

- [1]这是一封国际学术交流联系函件。学术交流联系信件的特点是开门见山,直奔主题;简明扼要,清晰准确;彬彬有礼,热情适度。
- [2]以上是对方的地址和称谓,正式或初次联系宜写出。
- [3]因为写信用的是有学校抬头的信纸,所以落款处只列写到系,不必再写学校名称。

7.3.4 Reference Translation

一封有关科技交流的联系函

卡尔加里大学

工学院

电气与计算机工程系

1994-12-12
XXX 教授
机电工程系
武汉食品工业学院
武汉 湖北 430023
中华人民共和国

亲爱的 XXX 教授:

谨此回复您 1994 年 12 月 20 日的来函。您在信中询问关于前来卡尔加里大学作为期 6 个月的访问学者以及和我们的研究小组共同开展研究工作的可能性。您还表示,您所有的旅费和生活费都将由贵国政府承担。

您大概已经知道,我们小组从事的研究领域是自适应控制、模糊逻辑控制及神经网络控制在电力系统的应用。如果您对在这些领域中做研究有兴趣的话,我将乐于接纳您作为期 6 个月的访问学者。

您应该知道,加拿大的生活费用是很高的,您应确保能有足够的经费。我们还向所有来我们这里的外国来访者建议在加拿大逗留期间办理健康保险。

如果您打算来参加我们正在进行的研究活动,我可以欢迎您任何时候前来我们的实验室。如果您有兴趣前来卡尔加里,请与加拿大驻北京大使馆联系办理签证事宜。

您忠诚的
XXX 教授
电气与计算机工程系

Main Reference Materials

- [1] 戴文进,杨植新. 电气工程及其自动化专业英语. 北京:电子工业出版社,2005.
- [2] A. K. S. Bhat. *A unified approach for the steady-state analysis of resonant converters*. IEEE Trans. Industrial Electronics, Vol. 38, No. 4, pp. 251-259, Aug. 1991.
- [3] A. K. S. Bhat. *Fixed frequency PWM series-parallel resonant converter*. IEEE Trans. Industry Applications, Vol. 28, No. 5, pp. 1002-1009, 1992.
- [4] Anonymous. *History of the Development of the ARM at the ACORN*, 2000.
- [5] B. K. Bose. *Adjustable speed AC drives—A technology status review*. Proc. IEEE, Vol. 70, pp. 116-135, Feb. 1982.
- [6] B. K. Bose. *Power Electronics and AC Drives*, Englewood Cliffs, N. J. : Prentice Hall, 1986.
- [7] Floyd, Thomas L. *Digital Fundamentals (HRD)*. Prentice Hall, 2005.
- [8] HIMA Paul Hildebrandt GmbH + Co KG. HIMA Worldwide, 2008.
- [9] J. D. Irwin. *Basic Engineering Circuit Analysis*, 5th ed. Upper Saddle River, N. J. : Prentice Hall, 1996. D. Kraus, *Circuit Analysis*, St. Paul: West Publishing, 1991.
- [10] Julian W. Gardner, Vijay K. Varadan, Osama O. Awadelkarim. *Microsensors, MEMS and Smart Devices*. Beijing: Tsinghua University Press, 2004.
- [11] Karl Johan Åström. *Control System Design*. 2002.
- [12] Katsuhiko Ogata. *Modern Control Engineering* (Fourth Edition). Beijing: Tsinghua University Press, 2006.
- [13] Kuo, Benjamin C. , Golnaraghi, Farid. *Automatic control systems*. Beijing: Higher Education Press, 2003.
- [14] N. Mohan and T. Undeland, *Power Electronics: Converters, Applications, and Design*, New York: John Wiley & Sons, 1995.
- [15] P. C. Sen, *Thyristor DC Drives*, New York: John Wiley, 1981.
- [16] Philips *Semiconductors*, 80C51 Family. 1997.
- [17] QT113 *Charge-Transfer Touch Sensor*. QUANTUN Research Group, 2004.
- [18] Reinhold Ludwig, Pavel Bretchko. *RF Circuit Design Theory and Application*. Prentice Hall, 1999.
- [19] R. H. Engelmann and W. H. Middelndorf. *Handbook of Electric Motors*, New York: Marcel Dekker, 1995.
- [20] Richard C. Dorf, Robert H. Bishop, *Modern Control System*. Beijing: Science Press, 2005.
- [21] Theodore J. Williams, Zhixin Yang. *Computer-based Automation for Flour Milling*. Purdue University, 1988.

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任 and 行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396; (010) 88258888

传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036